# A Model-based Approach for Combined Tracking and Resolution Enhancement of Faces in Low Resolution Video

Annika Kuhl, Tele Tan and Svetha Venkatesh
*Curtin University of Technology*
*Australia*

## 1. Introduction

Wide area surveillance situations require many sensors, thus making the use of high-resolution cameras prohibitive because of high costs and exponential growth in storage. Small and low cost CCTV cameras may produce poor quality video, and high-resolution CCD cameras in wide area surveillance can still yield low-resolution images of the object of interest, due to large distances from the camera. All these restrictions and limitations pose problems for subsequent tasks such as face recognition or vehicle registration plate recognition. Super-Resolution (SR) offers a way to increase the resolution of such images or videos and is well studied in the last decades (Farsiu et al., 2004; Park et al., 2003; Baker & Kanade 2002). However, most existing SR algorithms are not suitable for video sequences of faces because a face is a non-planar and non-rigid object, violating the underlying assumptions of many SR algorithms (Baker & Kanade, 1999).

A common SR algorithm is the super-resolution optical flow (Baker & Kanade, 1999). Each frame is interpolated to twice its size and optical flow is used to register previous and consecutive frames, which are then warped into a reference coordinate system. The super-resolved image is calculated as the average across these warped frames. However, the first step of interpolation introduces artificial random noise which is difficult to remove. Secondly, the optical flow is calculated between previous and consecutive frames preventing its use as an online stream processing algorithm. Also, accurate image registration requires precise motion estimation (Barreto et al., 2005) which in turn affects the quality of the super-resolved image as reported in (Zhao & Sawhney, 2002). Optical flow in general fails in low textured areas and causes problems in registering non-planar and non-rigid objects in particular. Recent techniques like (Gautama & van Hulle, 2002) calculate sub-pixel optical flow between several consecutive frames (with non-planar and non-rigid moving objects) however they are unable to estimate an accurate dense flow field, which is needed for accurate image warping.

Although solving all the issues in a general case is difficult, as the general problem of super-resolution is numerically ill-posed and computationally complex (Farsiu et al., 2004), we address a specific issue: Simultaneous tracking and increasing super-resolution of known object types, in our case faces, acquired by low resolution video. The use of an object-specific 3D mesh overcomes the issues with optic flow failures in low textured images. We avoid the

Source: State of the Art in Face Recognition, Book edited by: Dr. Mario I. Chacon M.,
ISBN -3-902613-42-4, pp. 250, January 2009, I-Tech, Vienna, Austria

use of interpolation, and use this 3D mesh to track, register and warp the object of interest. Using the 3D mask to estimate translation and rotation parameters between two frames is equivalent to calculating a dense sub-pixel accurate optical flow field and subsequent warping into a reference coordinate system. The 3D mesh is subdivided, such that each quad is smaller than a pixel when projected into the image, which makes super-resolution possible (Smelyanskiy et al., 2000). It also allows for sub-pixel accurate image registration and warping and in addition, such a fine mesh improves the tracking performance of low-resolution objects. Each quad then accumulates the average colour values across several registered images and a high resolution 3D model is created online during tracking. This approach differs from classical SR techniques as the resolution is increased at the model level rather than at the image level. Furthermore only the object of interest is tracked and super-resolved rather than the entire scene which reduces computation costs. Lastly, the use of a deformable mask mesh allows for tracking of non-rigid objects.

The novelty of our approach is the use of an object-specific 3D model within a combined tracking and super-resolution framework which means that the super-resolved image is created during tracking. The 3D mesh model allows for accurate tracking of non-planar and low-resolution objects, and unlike optical flow based super-resolution methods our methods does not need initial resolution increase by interpolation, thus results in less blurred images. The resulting high-resolution 3D models can be used for a number of applications such as generating the object under different views and different lighting conditions.

## 2. Background and related work

Super-Resolution methods increase the resolution of a single image or a whole video sequence and can be formally treated as single frame or multi-frame approaches using spatial or frequency information (Huang & Tsai, 1984; Borman & Stevenson, 1998).

The simplest way of increasing the resolution of a single image is by interpolation using techniques like nearest neighbour or spline interpolation. However interpolation alone is not able to recover high-frequency details of the image or video and is therefore not truly regarded as `formal' SR (Park et al., 2003). More complex methods model the image formation process as a linear system (Bascle et al., 1996)

$$Y = AX + z \tag{1}$$

where X and Y is the high and low-resolution image respectively. The degrading matrix A represents image warp, blur and image sampling; z models the uncertainties due to noise. Restoring the high-resolution image X involves inverting the imaging process but this is computationally complex and numerically ill-posed, even though different constraints have been proposed in the last years (Baker & Kanade 2002).

SR methods that reconstruct super-resolved images from video sequences apply the image formation process of Equation 1 to several frames (Farsiu et al. 2004) or use a Bayesian approach to estimate images of higher resolution (Baker & Kanade 2002). Super-Resolution optical flow is another approach for combining several frames of a video sequence. But according to (Baker & Kanade 1999) most existing super-resolution algorithms are not suitable for video sequences of non-planar and/or non-rigid objects.

Pre-requisite for increasing the resolution of video sequences by combining several frames are sub-pixel shifts between consecutive frames. Typical SR techniques assume that the camera is moving as the scene is recorded. But a moving camera results in motion blur which decreases the quality of the images. Even though special cameras are able to increase the resolution of motion blurred images (Agrawal & Raskar 2007), according to (Ben-Ezra et al. 2005) traditional cameras should avoid motion blur as much as possible. However, in surveillance applications cameras are generally fixed and the object of interest is moving. We will capitalise on this arrangement to define the tracking and super-resolution requirements. This way the amount of motion blur can be avoided or is reduced to a minimum depending on the speed of the object and the frame rate of the camera.

Combining several frames to increase resolution requires accurate motion estimation techniques (Barreto et al. 2005) to register and warp consecutive frames into a reference coordinate system. Accurate image registration is important and does affect the quality of the SR results as shown in (Zhao & Sawhney 2002). SR optical flow uses optical flow to register consecutive frames and typically comprises the following five main steps

1. *Image Interpolation* - interpolate each frame to twice its size
2. *Image Registration* - estimate the motion field between consecutive frames
3. *Image Warping* - warp images into a reference coordinate system
4. *Image Fusing* - fuse images using mean, median or robust mean
5. *Deblurring* - apply standard deconvolution algorithms to super-resolved image

While this approach is well suited to increase the resolution of images of rigid and planar scenes, image registration is more difficult for non-rigid and non-planar low-resolution objects that are more subjected to occlusion and lighting changes. The first step of the SR optical flow algorithm interpolates each frame to twice its resolution using standard interpolation techniques like nearest neighbour or bilinear. But interpolation cannot recover high-frequency details in images. In addition it introduces artificial random noise that is difficult to remove in the deblurring step. Image warping, the third step, also involves interpolation of pixels which introduces further noise.

A similar approach which obtains the super-resolved texture during tracking is proposed by (Dellaert et al. 1998). This method tracks planar objects and predicts the super-resolved texture using a Kalman filter. In (Yu & Bhanu 2006) SR optical flow is extended and planar patches are used to track different parts of the face individually to account for non-rigidity. The resolution of the face is increased for these different facial parts individually but again no three-dimensional object-specific mask mesh is used. The authors of (Smelyanskiy et al., 2000) use a Bayesian approach for high-resolution 3D surface construction of low-resolution images given a user provided 3D model. Synthetic images are rendered using the surface model, compared with the real low-resolution images and the difference is minimised.

Recent studies involve the use of special cameras to capture super-resolved video sequences. The so called jitter camera is used in (Ben-Ezra et al. 2005) and creates sub-pixel offsets between frames during recording. They also show that motion blur degrades the result of super-resolution algorithms, even if the motion blur itself is known, and should therefore be avoided. The authors of (Agrawal & Raskar 2007) use a special so called flutter shutter camera. This camera preserves the high frequencies by opening and closing the shutter frequently during exposure. A single camera is extended to capture super-resolved stereo images in (Gao & Ahuja 2006). Our work, however, uses images captures by a standard digital camera.

## 3. Method overview

The image formation process and basic outline of our approach is illustrated in Figure 1. The degrading matrix A in Equation 1 models image warp, blur due to the optical system and motion as well as the down-sampling process caused by the finite and discrete imaging chip. We neglect the optical blur and keep the motion blur down to a minimum.

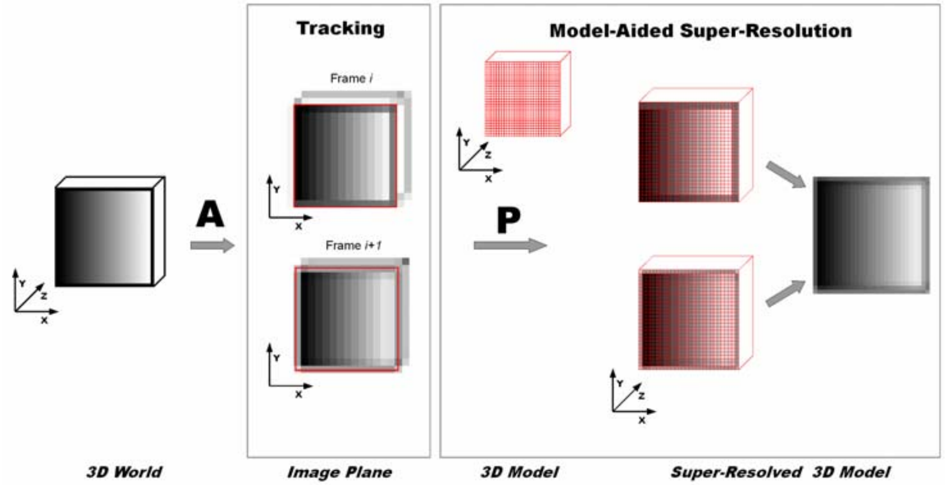


Fig. 1. Image formation process and basic outline of our approach. The matrix A degrades the images according to Equation 1. The 3D model of the object is textured by projecting every frame onto the model using the projection matrix P. The super-resolved texture is calculated as the mean across several frames.

During the image formation process the object is projected from the 3D world onto the image plane (Faugera 1993) and results, after being degraded by matrix A, in a finite number of discrete pixels as shown in Figure 1. The number of pixels that an object covers within the image depends on the size of the imaging chip, the optical lens, the size of the object itself and the distance between object and camera. As the camera or the object moves it may be projected onto different pixels of the imaging chip in different frames.

The black edges surrounding the gradient on the front side of the cube in Figure 1 are projected nearly exact into pixel centres resulting in 14 black pixels on either side of the cube in the image plane in frame *i*. A movement of the cube in front of the camera results in sub-pixel movements on the image plane. The black edges of the cube then may fall between pixels of the imaging chip resulting in grey edge pixels in frame *i+1* in Figure 1.

The tracking algorithm uses an object-specific deformable 3D model mesh to estimate translation and rotation parameters between consecutive frames which allows for accurate tracking of non-planar and non-rigid objects. Computer graphic techniques are used to subdivide the 3D model such that every quad of the mesh is smaller than a pixel when projected into the image, which makes super-resolution possible. By projecting the 3D model in several frames each quad accumulates different colour values over time. The super-resolved 3D model is then calculated as the mean colour value for each quad. Without loss of generality Figure 1 only shows the projection of one side of the planar cube; the same would be true for a non-planar and/or non-rigid object.

The super-resolved 3D model is created online during tracking and improves with every frame, whereas super-resolution optical flow incorporates consecutive and previous frames which prohibits its usage as an online stream processing algorithm. Furthermore, using an object-specific 3D model in a combined tracking and super-resolution approach inverses the image formation process in Equation 1. The subdivided 3D mesh represents the high-resolution object X that is down-sampled by projection into the image plane. The finer the mesh the higher the resolution of X and the larger the possible increase in resolution. Thus, interpolation, the first step of the optical flow algorithm, is unnecessary and the resulting super-resolved 3D model is less blurred by achieving the same resolution increase. This in turn makes deblurring (the last step of the optical flow algorithm) unnecessary. Lastly, using the 3D mesh for tracking equals image registration, warping and the estimation of a dense flow field, comprising steps 2 and 3 of the optical flow algorithm.

Furthermore we present an extended tracking approach that allows for the non-rigidity of objects during tracking. By incorporating a deformable mask mesh that allows for deformations during the tracking process super-resolution is possible despite non-planarity and non-rigidity.

## 4. Combined tracking and super-resolution

### 4.1 3D object tracking

A pre-requisite for building our super-resolved 3D model is the need to track the object in low-resolution video. We utilise an object-specific 3D mesh model which is manually fitted in the first frame. For fully automatic tracking, a combined appearance and geometric based approach similar to that in (Wen & Huang 2005) is used. We differ in that we use a subdivided fine mesh for the appearance approach and a constrained template matching algorithm for the geometric tracking as opposed to minimising a linear combination of action units. For each frame we apply both methods, i.e. the appearance based and the geometric based tracking. The one that results in the smallest tracking error will be used for the current frame. We describe each tracking approach in detail.

The **appearance based approach** is similar to the one in (Wen & Huang 2005). We differ in that we use a subdivided mask mesh which achieves better tracking results than a mesh that is coarser with respect to the pixel size. The warping templates $b_i$ are created from this subdivided mask as

$$b_i = I_0 - Q\left(\mathrm{P}\left(T_0 + n_i, X\right)\right) \text{ with } I_0 = \mathrm{P}\left(T_0, X\right) \tag{2}$$

where function P is the projection of 3D object points X to image coordinates using the initial transformation $T_0$; Q then maps RGB values to these coordinates. X is a vector containing the centre of gravity of each mask triangle, $n_i$ is the transformation parameter displacement and $I_0$ is a vector of the concatenated RGB-values of each projected triangle. The required intrinsic camera transformation parameters are obtained using camera calibration techniques as in (Zhang, 2000).

Objects are tracked by using the pose parameters of the previous frame as initialisation and solving for q and c for each frame i

$$I_0 - Q\left(\mathrm{P}\left(T_i^{app}, X\right)\right) \approx Bq + Uc \tag{3}$$

where the columns of B are the warping templates b$_i$. $T_i^{app}$ contains the transformation parameters for the appearance-based tracking at frame i and U are the illumination templates. Like in (Wen & Huang 2005) we use the first nine spherical harmonic bases, given the 3D mask mesh, for modelling lighting changes during tracking. Please refer to (Cascia et.al., 2000) for more details on the appearance based approach.

The **geometric based approach** uses a standard template matching approach that is restricted by the object-specific mask. We do not use action units to track facial movements like in (Wen & Huang 2005). Objects are tracked by projecting every vertex V of the mask into the previous image, using perspective projection. Around each projected vertex a rectangular template is cropped and matched with the current frame. The size of the patch is set to 1/6th of the whole object. Normalised cross-correlation is used to match this patch in the current frame within a window that is double the size of the template. In order to minimise the effect of outliers, the entire mask is fitted to the retrieved new vertex points v$_i$ in the current frame utilising the Levenberg-Marquardt algorithm

$$T_i^{geo} = \min_{T^{geo}} \sum_{j=1}^{l} \left( \mathrm{P}\left(T_i^{geo}, V_j\right) - v_j \right)^2 \tag{4}$$

where l is the number of mask vertices V and $T_i^{geo}$ contains the transformation parameters of the geometric-based approach at frame i. Again the transformation parameters are initialised with the previous frame.

During tracking each method is applied individually and a texture residual as the root mean squared error (RMSE) for the current frame i with respect to the first frame is calculated

$$RMSE(T_i) = \sqrt{\frac{1}{k} \sum_{j=1}^{k} \left( \mathrm{P}\left(T_0, X_j\right) - \mathrm{P}\left(T_i, X_j\right) \right)^2} \tag{5}$$

where k is the number of mask triangles X. The pose parameters T$_i$ of the method with the smallest RMSE will be used for the current frame i as

$$T_i = \min_{T_i} \left[ RMSE\left(T_i^{app}\right), RMSE\left(T_i^{geo}\right) \right] \tag{6}$$

where $RMSE\left(T_i^{app}\right)$ and $RMSE\left(T_i^{geo}\right)$ is the texture residual for frame i of the appearance-based and the geometric-based approach respectively. The tracking runs automatically once the mesh mask is manually fitted to the first frame of the sequence.

### 4.2 3D model-aided super-resolution

During tracking the resolution of the low-detailed object is gradually increased. To achieve this, every triangle of the object-specific mask is projected into the video using perspective projection. But in order to increase the resolution of the object, every triangle needs to be smaller than a pixel (Smelyanskiy et al., 2000).

As each mask triangle is projected into different frames of the sequence it is eventually assigned with different colour values for each frame as shown in Figure 1. Therefore the super-resolved mask I$_{SR}$ is calculated as the mean of the last k frames that result in an RMSE below a certain threshold ε

$$I_{SR} = \frac{1}{k} \sum_{i=1}^{k} \left( \mathrm{P}\left(T_i, X\right) \right) \text{ with } RMSE\left(T_i\right) < \varepsilon \tag{7}$$

Small tracking errors (RMSE) allow for an exact alignment of the 3D mask across frames whereas high RMSE result in blurring and distortion. The threshold ε depends on the initial object resolution. Low-resolution objects usually result in higher RMSE during tracking as image pixels are more likely to change due to the down-sampling process of the imaging chip. Furthermore the quality of the super-resolved mask $I_{SR}$ also depends on the total number of frames k used. But a larger number of frames increases the probability of introducing artificial noise as frames might not be aligned perfectly. The issue of choosing the appropriate number of frames k versus the quality of the super-resolved mask is experimentally evaluated in Section 4.4.

### 4.3 Extension to non-planar and non-rigid objects

In order to increase the resolution of various non-planar and non-rigid objects the tracking algorithm also needs to allow for deformations, i.e. the mask mesh representing the three-dimensional object needs to be deformable. This is especially an issue when tracking non-rigid objects like faces. We therefore propose an extended tracking and super-resolution algorithm and apply it to faces, as faces are a major field of interest especially in wide area surveillance systems.

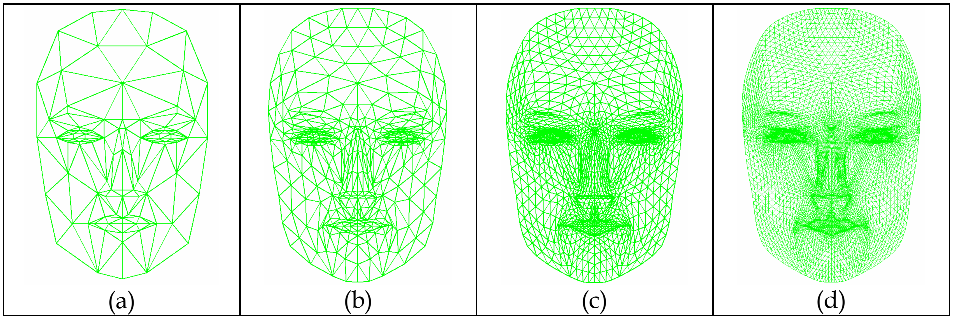


Fig. 2. (a) CANDIDE-3 face mask with 184 triangle, (b), (c) and (d) are subdivided masks after 1, 2 and 3 subdivision steps resulting in 736, 2944 and 11776 triangles. Source: (Kuhl et al.,2008) © 2008 IEEE

For tracking faces the CANDIDE-3 face model is used (Ahlberg, 2001). As shown in Figure 2 this triangular mesh consists of 104 vertices and 184 triangles and it is subdivided three times using the Modified Butterfly algorithm (Zorin & Schröder, 2000) to finally produce 5984 vertices and 11776 triangles. To allow for the non-rigidity of faces we use the CANDIDE-3 expression parameters for tracking mouth and eyebrow movements in low-detailed faces. More complex facial expressions often require a more detailed face model and high-resolution images, such as in (Goldenstein et al., 2003; Roussel & Gagalowicz, 2005; Wang et al., 2005).

The expression tracking is performed after the actual tracking for each frame. Using the expression parameters of the last frame the combined geometric and appearance based tracking approach is used to determine the position of the mesh model in the current frame. After that a global random search (Zhigljavsky, 1991) is performed to improve the RMSE around the mouth and the eyebrow region. A normal distribution with respect to the last expression parameters is used to sample 10 to 20 different values for each expression parameter. The parameter that results in the smallest RMSE is chosen for the current frame.

# 5. Experiments

## 5.1 Combined geometric and appearance-based 3D object tracking

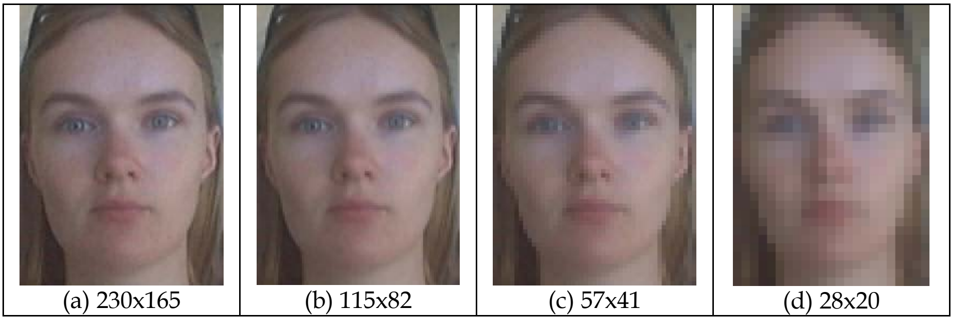
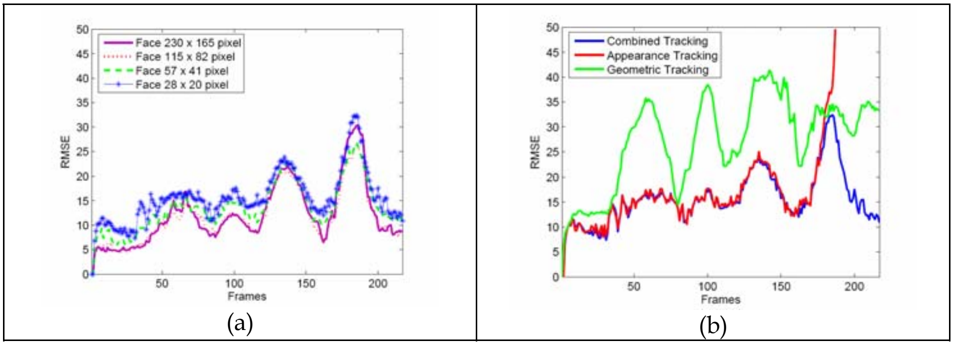

(a) 230x165 (b) 115x82 (c) 57x41 (d) 28x20

Fig. 3. Cropped faces for each face size used. Source: (Kuhl et al.,2008) © 2008 IEEE

In order to evaluate the tracking accuracy of the combined geometric and appearance based tracking algorithm, a video sequence of a face with translation and rotation movements is recorded at 15 frames per second and an initial resolution of 640x480 pixels. The face within one frame has an average size of 230x165 pixels. This resolution is divided into halves three times, resulting in face sizes of 115x82, 57x41 and 28x20 pixels with corresponding frame sizes of 320x240, 160x120 and 80x60 respectively. A cropped face for each face size used is shown in Figure 3.



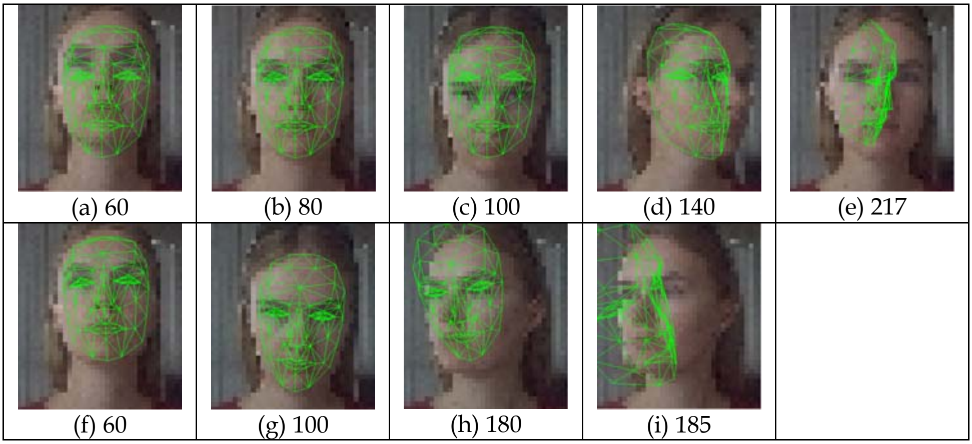
(a) (b)

Fig. 4. Tracking results for different size faces (a) and each tracking method applied individually to the video with the smallest resolution (b). Source: (Kuhl et al.,2008) © 2008 IEEE

For tracking faces we use the CANDIDE-3 face model (Ahlberg, 2001) as shown in Figure 2. In order to initialise this mask in the first frame of the sequence, we use the shape parameters of the CANDIDE-3 model to adjust the mask according to the persons individual face. After this initialisation, the mask is tracked automatically over more than 200 frames using the combined geometric and appearance-based approach as described in Section 4.1.

The results of the combined tracking algorithm applied to different face sizes is shown in Figure 4(a). The RMSE for measuring the tracking accuracy with respect to the first frame is defined in Equation 5. The variation in RMSE in frames 1 to 100 are due to translation and rotation around the horizontal x-axis, whereas the peaks at frames 130 and 175 respectively are mainly due to rotation around the vertical y-axis.

Faces between 230x165 and 115x82 result in similar RMSE, whereas faces with a resolution down to 57x41 result in a slightly increased tracking error, that is 22% larger on average. Even though the RMSE increased by about 41% when tracking faces with a resolution of 28x20, the algorithm is still able to qualitatively track the face to the end of the sequence.



(a) 60 (b) 80 (c) 100 (d) 140 (e) 217

(f) 60 (g) 100 (h) 180 (i) 185

Fig. 5. Sample frame shots of geometric based tracking (a)-(e) and appearance based tracking (f)-(i). Numbers denote different frames.

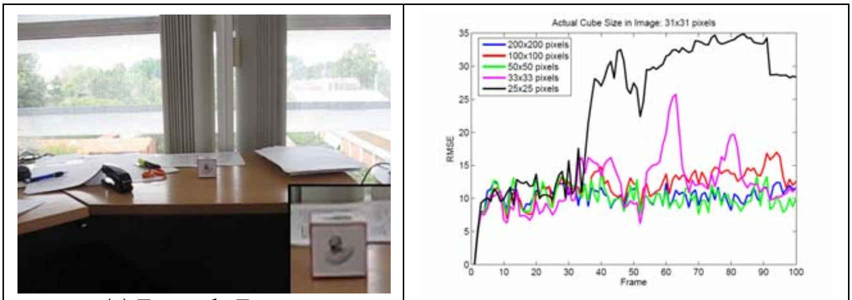
In comparison, Figure 4(b) shows the result of the geometric and appearance based tracking approach operating individually on the same video sequence, with the smallest face size of 28x20, as this face size is the most difficult to track. The geometric approach loses track just after 40 frames and as shown in Figure 5(a) this is due to small inter frame movements which causes the mask to stay in the initial position instead of following the face. The mask then recovers in frame 80 and loses track immediately afterwards as shown in Figure 5(b) and Figure 5(c). The geometric approach finally loses track around frame 140, from which it can not recover, as shown in Figure 5(d) and Figure 5(e). This shows that the geometric approach is not able to handle small inter frame movements due to image noise and low resolution.

The appearance based approach on the other hand results in small tracking errors from frame 1 through to frame 170 as shown in Figure 5(f) and Figure 5(g). But as the face turns the appearance approach loses track from which it cannot recover, as shown in Figure 5(h) and Figure 5(i). This is mainly due to large inter frame movements and the rotation of the face, resulting in partial occlution of the face.

Figure 4(b) shows that by combining the geometric and appearance based approach tracking is improved. Both approaches complement one another resulting in smaller RMSE than either of them individually. While the appearance based method tends to be more precise for small inter-frame movements, the geometric method is better for larger displacements. Furthermore the geometric approach applies template matching between the current and the previous frame, while the appearance approach is based on the comparison of the current frame with the first frame. Thus, the combination is more stable and precise and able to track even small size faces down to a size of 28x20 pixels.

### 5.2 Combined tracking and super-resolution

We tested the performance of the combined geometric and appearance-based tracking algorithm with different mask sizes. As described in Section 3, super-resolution is only possible when the mask mesh is subdivided such that every quad or triangle is smaller than a pixel when projected into the image. The following experiment evaluates the effect of the mesh size on the tracking performance.



(a) Example Frame

(b) Object Size 31x31

| | Object Size | |
|---|---|---|
| Mesh Size | 31x31 | 17x17 |
| 13x13 | - | 20.84 |
| 16x16 | - | 23.49 |
| 20x20 | - | 18.78 |
| 25x25 | 23.88 | 19.09 |
| 33x33 | 12.03 | 17.78 |
| 50x50 | 10.03 | 15.62 |
| 100x100 | 11.88 | 16.25 |
| 200x200 | 10.40 | - |

(c) Mean RMSE across 100 frames

Fig. 6. Mean tracking RMSE across 100 frames for a planar objects, the front side of a cube (a). Using an object mask mesh that is finer than the object results in smaller tracking errors.

The front side of a cube as shown in Figure 6(a) is tracked across 100 frames of size 640x480. This planar patch covers an area of 31x31 pixels within the image. For tracking this patch we used a 3D model mesh similar to the one in Figure 1. This mesh is equally subdivided into 25x25, 33x33, 50x50, 100x100 and 200x200 quads. Figure 6(b) shows the result of the combined tracking approach when different mesh sizes were used.

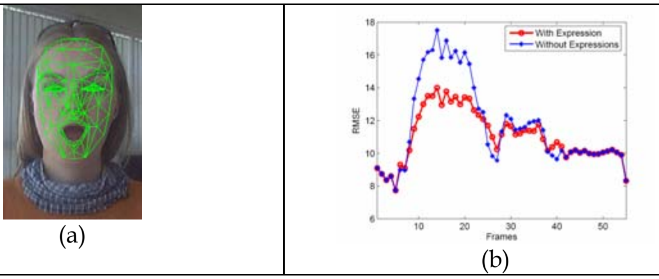
For tracking a planar patch of size 31x31, the best result is achieved with mesh sizes larger than 33x33, whereas further subdivision does not improve the tracking. Using a mesh that is coarser with respect to the pixel size loses track easily and results in higher RMSE during tracking as shown in Figure 6(b). Such a coarse mesh is a under representation of the object, resulting in higher tracking errors. By using a larger number of quads the mesh is able to account better for appearance changes due to sub-pixels movements.

Another video of the same object was recorded at greater distances, resulting in a cube of size 17x17 pixels. The mask mesh used for tracking consists of 13x13, 16x16, 20x20, 25x25, 33x33, 50x50 or 100x100 quads. The corresponding mean tracking errors are shown in Table 6(c). Again the best tracking results are achieved with a mask mesh that contains more quads than the pixels covered by the object within the image.

This shows that not only the super-resolution benefits from the tracking algorithm but also the super-resolution benefits tracking. The combined geometric and appearance-based tracking approach achieves best results when a fine model mesh is used. This fine mesh should ideally be subdivided such that every quad or triangle is smaller than a pixel when projected into the image. In practice a mesh that is double the size has proven to be the best trade-off between accuracy and speed.

### 5.3 Expression tracking

In order to evaluate the performance of the expression tracking approach, we recorded a video of a face with a resolution of 230x165 with mouth and eyebrow movements. One frame of this sequence is shown in Figure 7(a). The graph in Figure 7(b) compares the result of the expression tracking with the combined geometric and appearance based tracking approach without expression tracking. Frames 10 to 22 contain mouth openings and frames 28 to 38 contain eyebrow movements. The graph shows clearly that the expression tracking improves the result of the combined tracking approach by reducing the RMSE for each frame.



(a)

(b)

| Face Resolution | With Expression | Without Expression | Improvement |
|---|---|---|---|
| 210 x 145 | 10.89 | 11.50 | 5.3% |
| 105 x 73 | 10.96 | 11.60 | 5.5% |
| 52 x 37 | 13.92 | 14.51 | 4.1% |
| 26 x 18 | 16.23 | 16.40 | 1.0% |

(c)

Fig. 7. Results of the expression tracking, (b) compares the tracking RMSE with and without expression tracking, (c) shows the mean tracking RMSE of 56 frames for different face resolutions.

Furthermore we again cut the resolution of the video in halves three time resulting in face resolutions of 210x145, 105x73, 52x37 and 26x18 pixels in size. The mean tracking error across 56 frames for each resolution is shown in Table 7(c). While the difference in RMSE amounts to about 0.60 (between 5.5% and 4.1% improvement factor) for the first three resolution levels, a face resolution of 26x18 only results in a RMSE difference of 0.17 compared to the tracking approach without expressions. This equals an improvement factor of only 1.0%.

The smaller the resolution of the face, the larger the RMSE as shown in Figure 4(a) and Figure 7(c). A face that is captured in high-resolution, results in a large number of pixels representing this face. But the smaller the number of representative pixels, the more likely they are to change over time. Due to the discretisation of the imaging sensor certain face regions, e.g. the eyes, result in very few pixels being allocated to them. The colour value of these pixels is most likely to change over time as the camera or the face moves and these areas fall on or between different image pixels. Therefore, tracking expressions of low-resolution faces does not improve the overall RMSE significantly.

### 5.4 3D model-aided super-resolution

In order to increase the resolution of the object of interest the object-specific 3D model must be subdivided into a fine mesh. Each quad or triangle must me smaller than a pixel when projected into the image to make super-resolution possible as illustrated in Figure 1. The possible increase in resolution depends on the size of the 3D model mesh. The finer the mesh, the larger the possible increase in resolution but however more frames are needed.

The following experiments quantitatively evaluate the number of frames needed to achieve different resolutions. We therefore tracked a little cube across more than 200 frames of a video sequence recorded at a resolution of 320x240 with the cube of size 95x95. This video is sub-sampled three times resulting in resolutions of 160x120, 80x60, 40x30 and corresponding cube sizes of 48x48, 24x24, and 12x12 respectively. In order to diminish the effect of tracking errors the cube is tracked at the highest resolution of 95x95. The estimated pose parameters are then used for the cube of size 24x24 and 12x12. An example frame of both these resolutions is shown in Figure 8(a) and Figure 8(c) respectively. For comparison Figure 8(b) and Figure 8(d) show these frames after their resolution is doubled using bilinear interpolation.

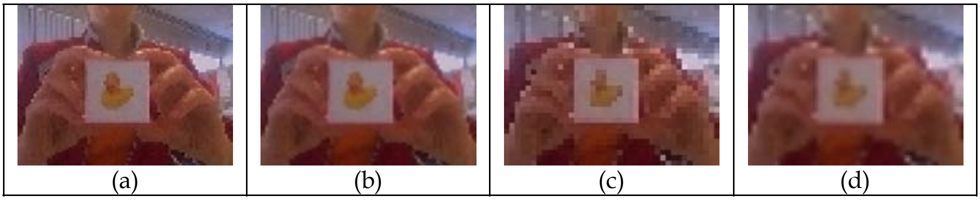

(a) (b) (c) (d)

Fig. 8. Example frame of resolution (a) 80x60 and (c) 40x30 with cube sizes of 24x24 and 12x12 respectively. Figure (b) and (d) show these frames after they have been doubled in size using bilinear interpolation.

For increasing the resolution of the cube of size 24x24 we used a 3D model as shown in Figure 1 and subdivided it into 25x25, 50x50 and 100x100 equal size quads. This mesh is projected into every frame of the sequence and the super-resolved 3D model $I_{SR}$ is created according to Equation 7 by combining 1, 10, 20, 50, 100 or 200 frames with the results shown in Figure 9.

Using a mesh of size 25x25 cannot increase the resolution of a cube of size 24x24. Although, calculating the mean across about 20 frames removes the noise of the camera and partially recovers the eyes of the duck that are not visible in the first frame as shown in Figure 9(c). Using a mesh that is double the size (50x50) results in a more detailed image but after about 20 to 50 frames the maximal possible resolution is achieved and adding more frames does not improve the resolution further as shown in Figure 9(i) and Figure 9(j) respectively.

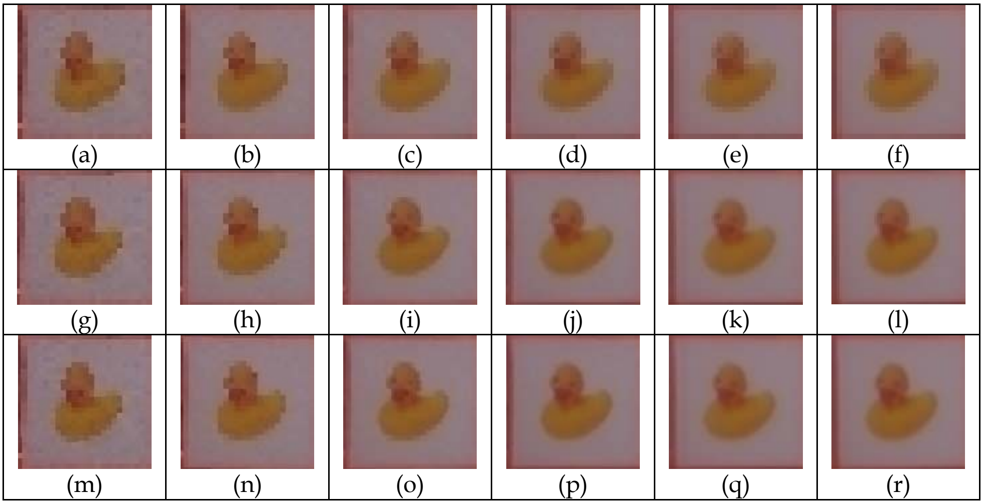


Fig. 9. Super-Resolution results of a cube of resolution of 24x24 pixels using mesh sizes of 25x25 (top), 50x50 (middle) and 100x100 (bottom) after 1, 10, 20, 50, 100 and 200 frames respectively.

Using a mesh of size 100x100 equals a possible resolution increase of four times in each dimension. But to achieve this increase at least 50 frames are needed as shown in Figure 9(p). But taking the mean across such a large number of frames also introduces noise as shown in Figure 9(r). This noise results from slightly misaligned frames and from the averaging process itself.

The same experiment is done for the cube of size 12x12 using mesh sizes of 20x20, 40x40 and 80x80. The result after combining 1, 10, 20, 50, 100 and 200 frames is shown in Figure 10. Using the mesh of size 20x20, which equals a resolution increase of 166%, requires about 20 to 50 frames (Figure 10(c) and Figure 10(d)). Adding more frames does not improve the result further.

Using a mesh of 40x40 results in a possible resolution increase of 3.3 times in each dimension and about 100 frames are needed to achieve this increase as shown in Figure 10(k). Trying to increase the resolution 6.6 times requires a mesh size of 80x80 and about 200 frames as shown in Figure 10(r). Even though the outer shape of the duck is recovered in greater detail, the overall result is very noisy and blurry as a result of taking the mean across 200 frames.

In practice it is therefore not recommended to increase the resolution of an object by more than 2 to 3 times. The higher the increase in resolution the higher the number of frames needed to achieve this resolution which in turn results in more noise. It is therefore a trade-off between possible resolution increase and number of frames. Furthermore the tracking error ε in Equation 7 also influences the resulting super-resolution image. Large tracking errors lead to a misalignment of frames resulting in noisy and blurred super-resolution images.

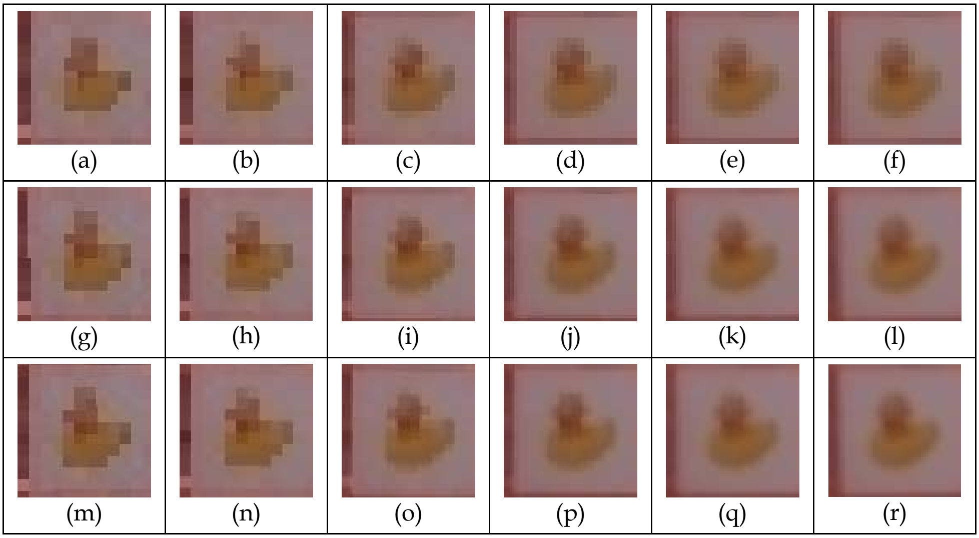


Fig. 10. Super-Resolution results of a cube of resolution of 12x12 pixels using mesh sizes of 20x20 (top), 40x40 (middle) and 80x80 (bottom) after 1, 10, 20, 50, 100 and 200 frames respectively.

Simple objects like a cube allow for an equal subdivision of the 3D model mesh. However, more complex objects may require a 3D model that consists of different size quads or triangles, thus resulting in a varying resolution increase across the mask mesh. For increasing the resolution of faces, the CANDIDE-3 mask as shown in Figure 2 is used. This mask is finely sampled around the eyes, the mouth and the nose region. These areas are also the most textured and the most important part of the face and therefore a finer sampled mask allows for a larger increase in resolution in these areas.

Evaluating the number of frames needed to achieve a certain resolution is more difficult using such a complex mask as the resolution increase varies across the entire mask due to different size triangles. In order to evaluate the minimum number of frames needed to create a face mask of higher resolution, faces are tracked in videos with minimal head movements in order to keep the tracking error to a minimum. The average sizes of the faces are 60x40 and 30x20 using videos of resolution 160x120 and 80x60 respectively. These resolutions are sub-sampled from the original video sequence recorded with 640x480. The mask mesh is manually fitted to the first frame of each sequence and then tracked fully automatically across more than 200 frames.

The super-resolved mask $I_{SR}$ is calculated using between 1 to 200 frames with the smallest tracking RMSE according to Equation 7. The mean tracking RMSE amounted to 10.9 and 14.9 for faces of size 60x40 and 30x20 respectively. We use the CANDIDE-3 mask that is subdivided three times as shown in Figure 2(d) for both face sizes. The high-resolution mask created from the 30x20 faces is then compared with a single frame of double the resolution (60x40) and the mask created from the 60x40 faces is compared with the face of 120x80 respectively. We use the mean colour difference $E_{colour}$ to compare two face masks $I_1$ and $I_2$ consisting of k triangles each

$$E_{colour} = \frac{1}{k} \sum_{i=1}^{k} \left| I_1(i) - I_2(i) \right|_2 \qquad (1)$$

Figure 11 shows the results for six different persons. Common to all persons and face sizes is the strong error decrease within the first 20 to 30 frames. Within the first 20 frames faces of size 60x40 increase resolution most significantly with respect to a face of double the size. Faces of size 30x20 are smaller and therefore more frames are needed to achieve the same resolution increase using a mask mesh with the same number of triangles. After about 20 to 30 frames the resolution increases most significantly. These results are comparable to the results of the cube shown in Figure 9 and Figure 10.

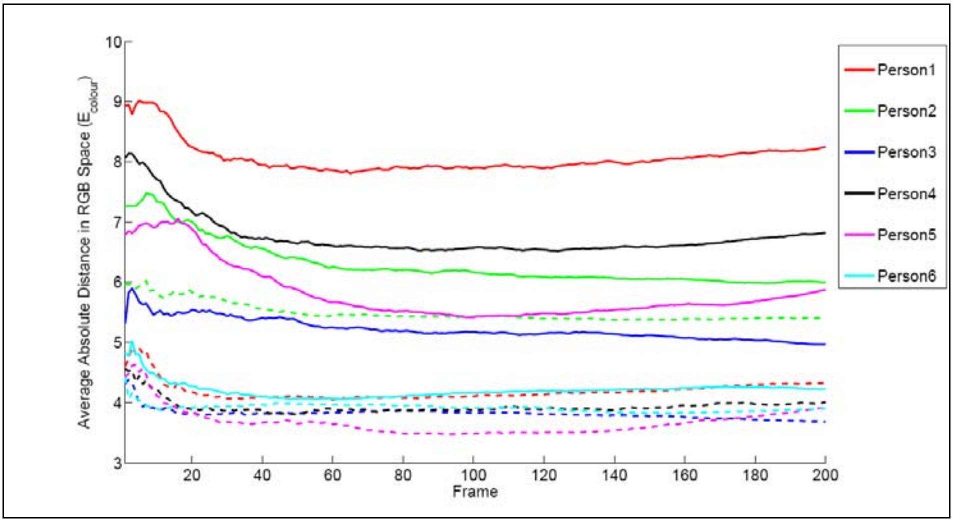


Fig. 11. Quality of the super-resolved 3D face mask using different resolutions (60x40, dotted line and 30x20 solid line) and number of frames (x-aches). Source: (Kuhl et al.,2008) © 2008 IEEE

For a qualitative comparison the super-resolution result of person 4 is shown in Figure 12. The faces on the left show the facial mask that is textured with a single frame of face size 60x40 (Figure 12(a)) and 30x20 (Figure 12(f)) respectively. The second, third and fourth collumn show the increase in resolution after 20 (Figure 12(b) and Figure 12(g)), 50 (Figure 12(c) and Figure 12(h)) and 200 (Figure 12(d) and Figure 12(i)) frames have been added to the super-resolved mask.

As the first step of the super-resolution optical flow algorithm is to double the size of the input images using interpolation techniques (see Section 2), we used bilinear interpolation to increase the size of the input video. The result after combing 200 frames of the interpolated input frames is shown in Figure 12(e) and Figure 12(j) for face sizes of 60x40 and 30x20 respectively. Even though the input images are doubled in size the resulting super-resolved faces show less detail and are more blurred. Interpolation does not recover high-frequencies and on the contrary introduces further noise; it should therefore be avoided during the super-resolution process.

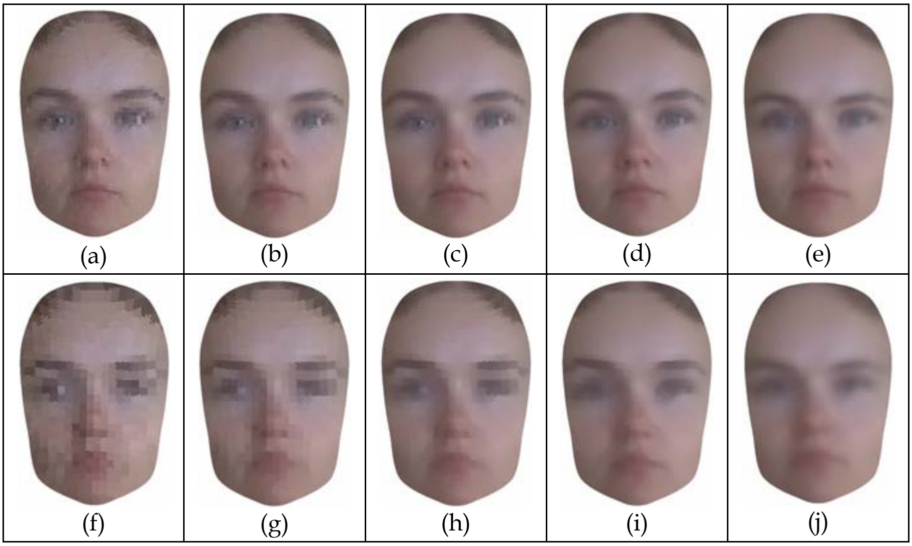


Fig. 12. Results of the combined tracking and super-resolution approach for face sizes 60x40 (top) and 30x20 (bottom) after 1, 20, 50 and 200 frames respectively. The last column shows the result after 200 frames when the input images are interpolated to double the size using bilinear interpolation. Source: (Kuhl et al.,2008) © 2008 IEEE

### 5.5 Comparison with super-resolution optical flow

We are comparing our approach with super-resolution optical flow by applying both algorithms to video sequences recorded in our lab as well as surveillance video of faces. Our implementation of the super-resolution optical flow follows the first four steps as outlined in Section 2. We abstained from using deconvolution techniques as this is an additional option for both our approach and the optical flow method to further increase the quality of the super-resolved images. The optical flow between consecutive frames is calculated using (Gautama & van Hulle, 2002) and the mean is used to calculate the super-resolved image.

At first we recorded a video sequence of a cube at 15 frames per second and with a resolution 320x240, where the cube is about 100x100 pixels in size. The proposed combined geometric and appearance based approach is used to track this cube across more than 100 frames resulting in a mean tracking error of 9.16. The pose parameters are then used to project the 3D model into every image and k frames with the smallest tracking error are used to calculate the super-resolved image $I_{SR}$ according to Equation 7.

The 3D model of the cube is subdivided into 400x400 quads before projected into the image which, under ideal conditions, equals a resolution increase of 400%. The result is shown in Figure 13. After about 20 frames (Figure 13(d) and Figure 13(j)) the maximum resolution increase is reached and further added frames result in increased blur due to tracking errors. For comparison we doubled the size of each frame using bilinear interpolation before creating the super-resolution image. Again a cube with 400x400 quads is used and the result is shown in Figure 13(m) to Figure 13(r). Even though the input images are doubled in size the resulting super-resolution images after about 20 frames (Figure 13(p)) do not show a significant resolution increase compared to the one without initial interpolation (Figure

13(j)). On the contrary, like in Figure 12, the resulting super-resolution images show a greater amount of blur as interpolation cannot recover high-frequency details.

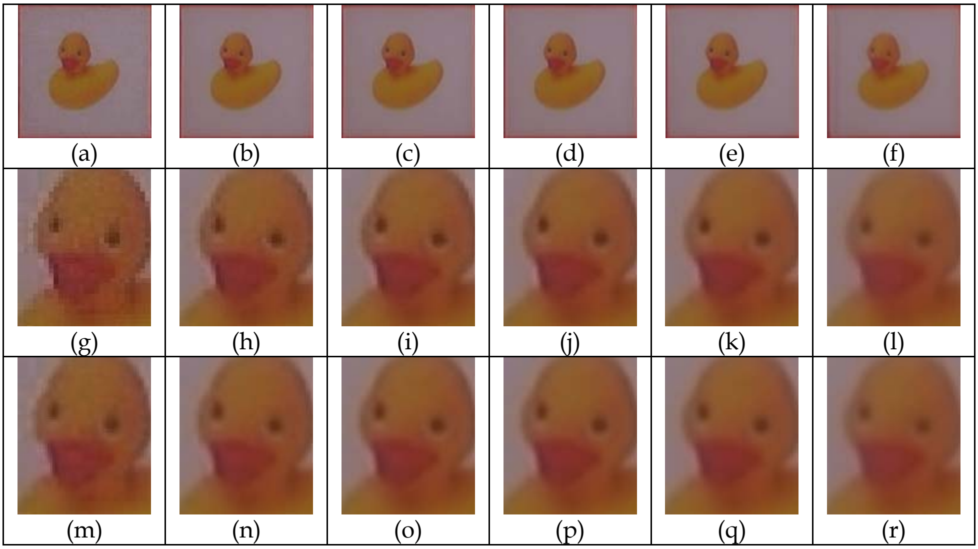


Fig. 13. Result of the proposed combined tracking and super-resolution approach after combining (a) 1, (b) 5, (c) 10, (d) 20, (e) 50, (f) 100 frames. Figure (g)-(l) show cropped parts of every Figure (a)-(f) respectively and Figure (m)-(r) show the result after each input image was doubled in size using bilinear interpolation.

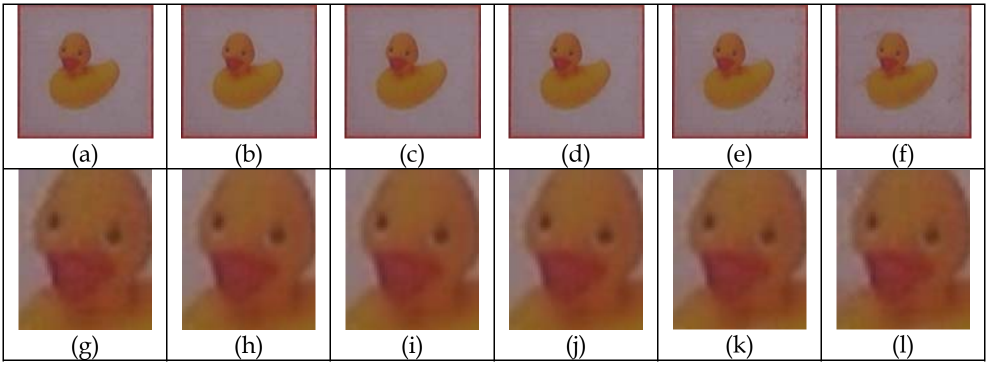


Fig. 14. Result of the super-resolution optical flow algorithm after taking the mean across (a) 1, (b) 5, (c) 10, (d) 20, (e) 50 and (f) 100 frames. A cropped part of each image is shown in (g)-(l) respectively.

Super-resolution optical flow increase the resolution of each frame by interpolation, combining several frames afterwards enhances the quality of the super-resolved image but does not further increase the resolution. The increase in resolution is usually fixed at 200%. This is contrary to our approach as we avoid interpolation to allow for less blurred images. But optical flow based methods require less frames as shown in Figure 14. The quality of the optical flow super-resolved image increases most significantly within the first 5 to 10 frames as shown in Figure 14(b) and Figure 14(c). This corresponds to the number of frames most optical flow based super-resolution methods use, like in (Baker & Kanade, 1999). The addition of more frames results in more blurred and noisy images as estimating an accurate dense flow field across a large number of frames is difficult and erroneous especially in low-resolution images. This is clearly visible in Figure 14(e) and Figure 14(f). Estimating a dense optical flow field across 50 to 100 frames is erroneous and the calculation of the mean across such a large number of frames results in artefacts and distortion.

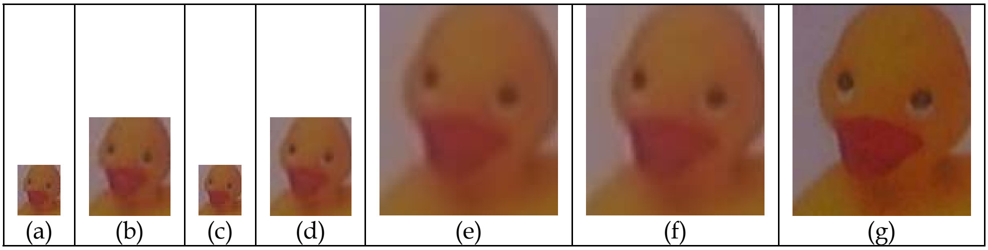


Fig. 15. (a) Cropped original frame with a cube of 100x100 pixels, (b) result after applying bilinear interpolation, (c) result of the optical flow after 10 frames without interpolation, (d) result of the optical flow after 10 frames with interpolation, (e) result of our approach after 20 frames with interpolation (f) result of our approach after 20 frames without interpolation, and (g) cropped cube of size 400x400 pixels.

Figure 15 summarises the results of both methods. The original video is recorded at a resolution of 320x240 and the cube is of size 100x100. A single cropped frame is shown in Figure 15(a). The simplest way of increasing the resolution is by interpolation, the result of bilinear interpolation is therefore shown in Figure 15(b). First we applied the super-resolution optical flow algorithm without an initial resolution increase by interpolation. The result after taking the mean across 10 frames is shown in Figure 15(c). The quality of the image is improved, i.e. the image noise is reduced, but the resolution remained unchanged. Interpolation is needed to increase the resolution, thus Figure 15(d) shows the result of the optical flow algorithm using bilinear interpolation to double the resolution of the input frames.

We also applied our approach to the video sequence that has been doubled in size using interpolation. The resulting super-resolved image, as shown in Figure 15(e), is more blur red and shows less detail, as a result of the interpolation, compared to the super-resolved image in Figure 15(f) that is calculated from the original input sequence. Furthermore the subdivision of the 3D model into a fine mesh allows for a greater increase in resolution compared to optical flow based methods. Lastly, Figure 15(g) shows a cropped image of the cube of size 400x400 pixels, which equals a resolution increase of 400% compared to Figure 15(a).

Furthermore, we tested our approach on non-planar and non-rigid objects, in this case that of surveillance video of people entering a bus. The video is recorded with a resolution of 640x480 at 23 frames per second due to dropped frames. Each frame is sub-sampled to half the resolution resulting in 320x240 pixels. The face within one frame is about 32x25 pixels, a single cropped frame of two different persons is shown in Figure 16(a) and Figure 16(f). We applied the combined geometric and appearance based approach to track these faces across about 40 frames.

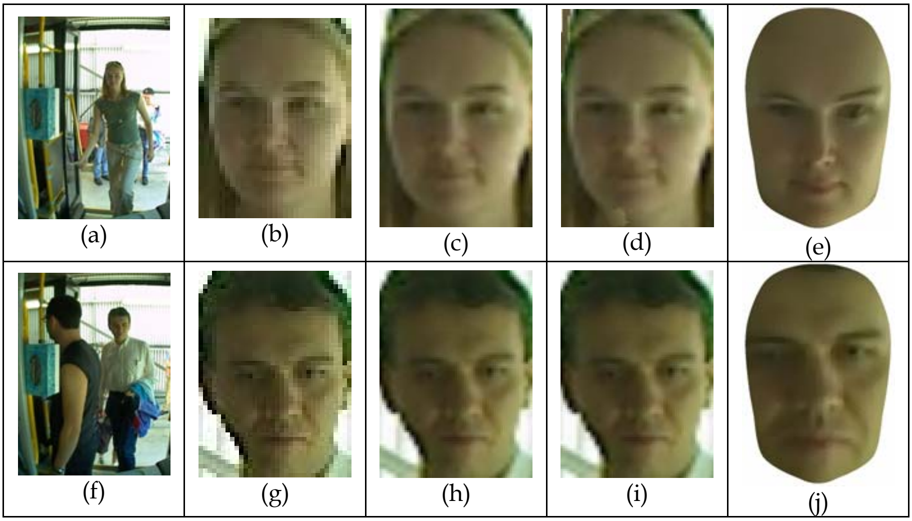


Fig. 16. Original image (a) and (f), cropped face (b) and (g), bilinear interpolation (c) and (h), optical flow result combining 20 frames (d) and (i) and our approach combining 20 frames (e) and (j).

We omitted the use of the extended expression tracking approach because we assume that people have a fairly neutral expression when entering the bus. Furthermore the low resolution of the face does not justify the runtime overhead of the proposed expression tracking. Also the difference in tracking error when tracking with and without expressions decreases with the resolution as shown in Section 4.3. If expressions occur during tracking nonetheless the tracking error will increase. But as the super-resolved image is created by using only frames below the threshold ε (Equation 7), frames with expressions will not be used and therefore not affect the result.

Optical flow is feasible for tracking planar objects like the front side of a cube in the last experiment but tracking non-planar and non-rigid objects like faces poses more challenges, especially when trying to estimate a dense flow field across a large number of frames. The result of the optical flow combining 20 frames is shown in Figure 16(d) and Figure 16(i). The chin area in Figure 16(d) shows slight distortion due to erroneous flow vectors. Figure 16(b) and Figure 16(g) show the initial frames after they have been doubled in size using bilinear interpolation.

A comparative result of the combined tracking and super-resolution method after 20 frames is shown in Figure 16(e) and Figure 16(j). The super-resolved faces of our approach are less blurred and shows more detail compared to the result of the optical flow. Again the initial interpolation introduces artificial random noise, whereas our approach does not need an initial interpolation; but the achieved resolution increase after 20 frames is equal or slightly higher.

Another advantage of our approach is the further use of the created super-resolved 3D model. In case of faces these models can be used to generate various face images under different pose and lighting in order to improve the training of classifiers or the super-resolved 3D model itself can be used for 3D face recognition.

## 6. Conclusion

We proposed a combined tracking and super-resolution algorithm that increases the resolution online during the tracking process. An object-specific 3D mask mesh is used to track non-planar and non-rigid objects. This mask mesh is then subdivided such that every quad is smaller then a pixel when projected into the image. This makes super-resolution possible and in addition improves tracking performances. Our approach varies from traditional super-resolution as the resolution is increased on mask level and only for the object of interest rather than on image level and for the entire scene.

We demonstrated our combined geometric and appearance based tracking approach on sequences of different size faces and showed that our approach is able to track faces down to 28x20 pixels in size. The combination of these two tracking algorithms achieves better results than each method alone.

The tracking algorithm is further extended to allow for the non-rigidity of objects. We applied this to faces and expressions. Experiments showed that the proposed method for expression tracking reduces the mean tracking error and thus allows for a better alignment of consecutive frames, which is needed to create super-resolution images.

The proposed 3D model-aided super-resolution allows for a high increase in resolution; the finer the 3D mesh the higher the possible increase in resolution. Therefore we experimentally estimated the number of frames needed to achieve a certain resolution increase. In practice about 20 to 30 frames are needed to double the resolution. Increasing the resolution further is limited by the number of frames used as well as the tracking error. Large tracking errors and the averaging process across a large number of frames introduces noise that decreases the quality of the super-resolved image.

We demonstrated our method on low resolution video of faces that are acquired both in the lab and in a real surveillance situation. We show that our method outperforms the optical flow based method, and performs consistently better for longer tracking durations in video that contain non-planar and non-rigid low-resolution objects. The combined tracking and super-resolution algorithm increases the resolution on mask level and makes interpolation, the first step of the optical flow algorithm, redundant. The resulting super-resolved 3D model is less blurred by achieving the same or a higher resolution increase. This in turn makes deblurring, the last step of the optical flow algorithm, unnecessary. Furthermore the super-resolved 3D model is created online during tracking and improves with every frame, whereas super-resolution optical flow incorporates consecutive and previous frames which prohibits its usage as an online stream processing algorithm.

## 7. References

Agrawal, A. & Raskar, R.. (2007). Resolving Objects at Higher Resolution from a Single Motion-blurred Image. *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-8, June 2007.

Ahlberg, J. (2001). CANDIDE-3 - An updated parameterized face. *Technical Report LiTH-ISY-R-2326*, Department of Electrical Engineering, Linkping University, Sweden.

Baker, S. & Kanade, T. (1999). Super-resolution optical flow. *Technical Report CMU-RI-TR-99-36*, Robotics Institute, Carnegie Mellon University, October, 1999.

Baker, S. & Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (9), 1167 - 1183.

Barreto, D.; Alvarez, L. D. & Abad, J. (2005). Motion Estimation Techniques in Super-Resolution Image Reconstruction. A Performance Evaluation. *Virtual Observatory: Plate Content Digitization, Archive Mining and Image Sequence Processing, iAstro workshop*, Sofia, Bulgaria, p. 254-268

Bascle, B.; Blake, A. & Zisserman, A. (1996). Motion deblurring and super-resolution from an image sequence. *Proceedings of the 4th European Conference on Computer Vision*, Volume II. Springer-Verlag, London, UK, pp. 573-582.

Ben-Ezra, M.; Zomet, A. & Nayar, S. K. (2005). Video super-resolution using controlled subpixel detector shifts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (6), 977-987.

Borman, S. & Stevenson, R. L. (1998). Super-resolution from image sequences - A review. *Proceedings of the 1998 Midwest Symposium on Systems and Circuits*. IEEE Computer Society, Washington, DC, USA, pp. 374-378, August 1998.

Cascia, M. L.; Sclaroff, S. & Athitsos, V. (2000). Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Computer Society*, 2000, 22, 322-336

Chiang, M. & Boult, T. (2000). Efficient super-resolution via image warping. *Image and Vision Computing*, Elsevier, Vol. 18 (10), July 2000, pp. 761-771.

Dellaert, F.; Thorpe, C. & Thrun, S. (1998). Super-resolved texture tracking of planar surface patches. *IEEE/RSJ International Conference on Intelligent Robotic Systems*, October 1998, pp. 197-203.

Farsiu, S.; Robinson, D.; Elad, M. & Milanfar, P. (2004). Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology* 14 (2), Wiley, August 2004, 47-57.

Faugera, O. (1993). Three-Dimensional Computer Vision - A Geometric View-point. *MIT Press*, Cambridge, MA, USA.

Gao, C. & Ahuja, N. (2006). A refractive camera for acquiring stereo and super-resolution images. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2), pp. 2316-2323.

Gautama, T. & van Hulle, M. (2002). A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Transactions on Neural Networks* 13 (5), September 2002, pp. 1127-1136.

Goldenstein, S. K.; Vogler, C. & Metaxas, D. (2003). Statistical cue integration in DAG deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (7)*, pp. 801-813, July 2003.

Huang, T. S. & Tsai, R. Y. (1984). Multi-frame image restoration and registration. *Advances in Computer Vision and Image Processing: Image Reconstruction from Incomplete Observations*, Thomas S. Huang (Ed.), London, 1984, Vol. 1, pp. 317-339, JAI Press.

Kuhl, A.; Tan, T. & Venkatesh, S. (2008). Model-based combined tracking and resolution enhancement. *Proceedings of the 2008 IEEE International Conference on Pattern Recognition*.

Lin, Z. & Shum, H.-Y. (2004). Fundamental limits of reconstruction-based super-resolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26 (1), pp. 83-97, Jan 2004

Park, S. C.; Park, M. K. & Kang, M. G. (2003). Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine* 20 (3), May 2003, pp. 21-36.

Roussel, R. & Gagalowicz, A. (2005). A hierarchical face behavior model for a 3D face tracking without markers. *Computer Analysis of Images and Patterns*. Vol. 3691. Springer, pp. 854-861.

Smelyanskiy, V.; Cheeseman, P.; Maluf, D. & Morris, R. (2000). Bayesian super-resolved surface reconstruction from images. *Proceedings. IEEE Conference on Computer Vision and Pattern Recognition*, vol.1, pp.375-382

Tanaka, M. & Okutomi, M. (2005). Theoretical analysis on reconstruction-based super-resolution for an arbitrary PSF. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE Computer Society, Washington, DC, USA, pp. 947-954, June 2005.

Wang, Y.; Gupta, M.; Zhang, S.; Wang, S.; Gu, X.; Samaras, D. & Huang, P. (2005). High resolution tracking of non-rigid 3D motion of densely sampled data using harmonic maps. *Tenth IEEE International Conference on Computer Vision*, pp. 388-395, Oct. 2005

Wen, Z. & Huang, T. (2005). Enhanced 3D geometric-model-based face tracking in low resolution with appearance model. *IEEE International Conference on Image Processing*, vol.2, pp. II-350-3, September 2005

Yu, J. & Bhanu, B. (2006). Super-resolution restoration of facial images in video. *Proceedings of the 18th International Conference on Pattern Recognition*. IEEE Computer Society, Washington, DC, USA, pp. 342-345.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11), pp. 1330-1334, Nov 2000

Zhao, W. & Sawhney, H. S. (2002). Is super-resolution with optical flow feasible?, *Proceedings of the 7th European Conference on Computer Vision-Part I.* Springer-Verlag, London, UK, pp. 599-613.

Zhigljavsky, A. A. (1991). Theory of global random search. *Dordrecht Netherlands : Kluwer Academic Publishers.*

Zorin, D. & Schröder, P. (2000). Subdivision for modeling and animation. *ACM Siggraph Course Notes*