This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044587

Reproduced with the kind permissions of the copyright owner.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright : 2008, IEEE
Model-based Combined Tracking and Resolution Enhancement

Annika Kuhl, Tele Tan, Svetha Venkatesh
Department of Computing, Curtin University of Technology
GPO Box U1987, Perth, 6845, Western Australia
a.kuhl@postgrad.curtin.edu.au, (t.tan, s.venkatesh)@curtin.edu.au

Abstract
Wide area surveillance requires high-resolution images of the object of interest derived possibly from only low-resolution video of the whole scene. We propose a combined tracking and resolution enhancement approach that increases the resolution of the object of interest during tracking. The key idea is the use of an object-specific 3D mesh model with which we are able to track non-planar objects across a large number of frames. This model is subdivided such that every triangle is smaller than a pixel when projected into the image to facilitate super-resolution on the model rather than on the image. We apply our approach to faces and show that it outperforms interpolation methods by achieving resolution enhancement, while being less blurred.

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 videos and is well studied in the last decades [6, 7, 3]. 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 [2].

Super-Resolution methods can either be applied to single images to several frames [6]. The authors of [3] use a Bayesian approach to estimate images of higher resolution. Super-Resolution optical flow [2] is proposed by combining several frames of a video sequence. But according to [2] most existing approaches are not suitable for non-planar and/or non-rigid objects.

An approach which obtains the super-resolved texture during tracking is proposed by [5]. They track planar objects and predicts the super-resolved texture using a Kalman filter. In [10], SR optical flow is used to track planar patches of 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 no 3D object-specific mesh is used. The authors of [8] 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.

We propose a resolution enhancement technique using an object-specific 3D mesh model for alignment. A combined geometric and appearance based tracking method uses this mesh to track low detail objects in video. The 3D mesh model is subdivided, such that each triangle is smaller than a pixel when projected into the image, which makes super-resolution possible [8]. Each triangle then accumulates the average colour values across several registered images and a higher resolution 3D appearance model of the object results. 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 reducing computation costs.

The novelty of our approach is the use of an object-specific 3D mesh 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 super-resolved 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 Method

We use a combined appearance and geometric tracking algorithm that makes use of an object-specific 3D mesh model to track objects in low-resolution videos. Throughout this paper we use a 3D mesh model for faces; note that any other non-planar 3D objects can be utilised as well. This mask is subdivided into a fine mesh, such that every triangle is smaller than a pixel. By projecting this mesh onto several frames, each triangle of the mesh accumulates different colour values over time. The super-resolved 3D model is then calculated as the mean colour value for each triangle.

2.1 3D Object Tracking

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 [9] 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. During the tracking process each method is applied individually and the best one is chosen for each frame.

Our appearance based approach uses the subdivided mask for calculating the warping templates \( b_i \) given as

\[
b_i = I_0 - Q(P(T_0 + n_i, X)) , \quad I_0 = P(T_0, X)
\]

where function \( P \) is the projection of 3D object points \( X \) to image coordinates using the initial transformation \( T_0 \) [9]. \( Q \) 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 standard camera calibration [11].

Objects are tracked by solving for \( q \) for each frame \( i \)

\[
I_0 - Q(P(T_i^{app}, X)) \approx Bq,
\]

where the columns of \( B \) are the warping templates \( b_i \). \( T_i^{app} \) contains the transformation parameters for the appearance-based (app) tracking at frame \( i \). Please refer to [4] for more details on the appearance approach.

The geometric approach projects all vertices \( V \) of the mask into the previous image. Small templates around each vertex are then matched with the current frame using normalised cross-correlation resulting in \( v_j \). The size of the templates is set to \( \frac{1}{4} \)th of the whole object. In order to minimise the effect of outliers, the entire mask is fitted to these retrieved vertex points \( v_j \) in the current frame utilising Levenberg-Marquardt

\[
T_i^{geo} = \min_{T_i^{geo}} \sum_{j=1}^{l} (P(T_i^{geo}, V_j) - v_j)^2, \quad (3)
\]

where \( l \) is the number of mask vertices \( V \) and \( T_i^{geo} \) contains the transformation parameters of the geometric-based (geo) approach at frame \( i \).

A texture residual is calculated as the root mean squared error (RMSE) for the current frame \( i \) with respect to the first frame

\[
RMSE(T_i) = \sqrt{ \frac{1}{k} \sum_{j=1}^{k} (P(T_0, X_j) - P(T_i, X_j))^2 }, \quad (4)
\]

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

\[
T_i = \min_{T_i} [RMSE(T_i^{app}), RMSE(T_i^{geo})], \quad (5)
\]

where \( RMSE(T_i^{app}) \) and \( RMSE(T_i^{geo}) \) is the texture residual for frame \( i \) of the appearance-based and the geometric-based approach respectively.

2.2 Model-Based Super-Resolution

After the low-resolution objects have been tracked, the computed pose parameters \( T_{1:N} \) for all \( N \) frames provide the information to align the facial mask to each frame. Each mask triangle is projected into every image of the sequence and will eventually be assigned with a different colour value for each image. This is due to the fact that every triangle is smaller than a pixel.

The super-resolved mask \( I_{SR} \) is then calculated as the mean colour value of \( n \) frames with the smallest tracking RMSE (Eq. 4)

\[
I_{SR} = \frac{1}{n} \sum_{i=1}^{n} (P(T_i, X)), \quad (6)
\]

Large RMSE indicate a bad alignment and may result in blurred and distorted super-resolved images \( I_{SR} \).

3 Experiments

3.1 3D Object Tracking

We utilise the CANDIDE-3 face model [1] to evaluate the accuracy of our tracking algorithm; that is a generic 3D face model that can be fitted to an arbitrary face by adjusting a number of shape parameters. This face mask is subdivided three times to finally produce 5984 vertices and 11776 triangles using the Modified Butterfly algorithm [12]. Fig. 1 shows the result of this
Figure 1. (a) Original CANDIDE-3 face mask, (b), (c) and (d) are subdivided masks after 1, 2 and 3 steps.

subdivision, which allows for increased resolution after tracking since every triangle is smaller than a pixel.

A video of a face with translation and rotation movements is recorded at 15 frames per second and a resolution of 640×480 pixels. The face within one frame has an average size of 230×165 pixels. This resolution is divided into halves three times, resulting in face sizes of 115×82, 57×41 and 28×20 pixels with corresponding frame sizes of 320×240, 160×120 and 80×60 respectively. A cropped face for each size is shown in Fig. 2.

Figure 2. Cropped face for each size used.

For each face size, the mask is initialised manually in the first frame using 14 shape parameters and then automatically tracked over more than 200 frames using the combined geometric and appearance-based approach as described in Sec. 2.1. Fig. 3 shows the results of the tracking algorithm applied to different face sizes.

Figure 3. Results for different resolutions.

The RMSE for measuring the tracking accuracy with respect to the first frame is defined in Eq. 4. 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 230×165 and 115×82 result in similar RMSE, whereas faces with a resolution down to 57×41 result in a tracking error, that is 22% larger on average. Even though the RMSE increased by about 41% when tracking faces with a resolution of 28×20, the algorithm is still able to qualitatively track the face through to the end of the sequence. Thus, this shows that the combined geometric and appearance-based tracking algorithm is well suited for tracking faces that are quite small in size. Best tracking results are achieved with a mask that has been subdivided three times, such that every triangle is smaller than a pixel.

In comparison, the appearance based approach loses track after 170 frames, when applied individually on the smallest image resolution (80×60). The geometric based approach alone results in a mean tracking error that is 75 % larger on average compared to the combined approach, as the mask loses track frequently.

3.2 Super-Resolution

After tracking, the super-resolved 3D mesh model is calculated as the mean of n frames with the smallest RMSE using Eq. 6. In order to evaluate the minimum number of frames needed to create a super-resolved mask, faces are tracked in videos with minimal head movements in order to keep the tracking error small. The average size of the faces are 60×40 and 30×20 using a resolution of 160×120 and 80×60 respectively. They were created by down-sampling a high-resolution video of size 320×240.

The super-resolved mask was calculated using between 1 to 200 frames with the smallest tracking RMSE. The mean tracking RMSE amounted to 10.9 and 14.9 for faces of size 60×40 and 30×20 respectively. The high-resolution masks created from the 30×20 and the 60×40 faces is then compared with the ground truth, which is a mask that is manually fitted to an image of double the resolution, 60×40 and 120×80 respectively.

We use the mean colour difference for comparison

$$E_{colour}^f = \frac{1}{k} \sum_{i=1}^{k} |I_f(i) - I_{SR}(i)|_2$$  \hspace{1cm} (7)$$

where $I_{SR}$ is the super-resolved mask (Eq. 6), $I_f$ is the mask at frame $f$ and $k$ is the number of mask triangles.

Fig. 4 shows the results for six different persons. Common to all persons and face sizes is the strong error decrease within the first 20 frames. The faces of size 60×40 (solid line) need just under 20 frames while 30×20 size faces (dotted line) need about 20 to 30 frames to increase resolution most significantly. The
$E_{\text{colour}}$ also depends on the correct alignment of the masks between different resolution levels which results in varying errors $E_{\text{colour}}$.

![Figure 4](image4.png)

**Figure 4.** Frames vs. quality of super-resolved mask.

The super-resolution result of Person4 is shown in Fig. 5. The faces on the left show the facial mask that is textured with a single frame with face sizes of $60 \times 40$ (5(a)) and $30 \times 20$ (5(f)) respectively. The second, third and fourth mask show the increase in resolution after 20 (5(b) and 5(g)), 50 (5(c) and 5(h)) and 200 (5(d) and 5(i)) frames have been added to the super-resolved mask. The last column compares these results with a single frame that has been increased to double its size using B-Spline interpolation. In both cases, in Fig. 5(e) and Fig. 5(j), the interpolated result is more blurred and less detailed.

![Figure 5](image5.png)

**Figure 5.** Resolution increase for face sizes $60 \times 40$ (top) and $30 \times 20$ (bottom).

This experiment shows that the resolution increases within the first 20 to 30 frames depending on the initial face size. As more frames are added to the mean face, the more blurred it becomes resulting from increasing tracking errors, misaligning the tracked mask. However, by choosing only frames with low RMSE (Eq. 4), the amount of blur is reduced to a minimum.

Compared to super-resolution optical flow our approach does not require initial interpolation and thus results in less blurred images. Furthermore, optical flow is not well suited for estimating a dense flow field of non-planar and non-rigid objects across a large number of frames which is needed for best results [2]. Space constraints prohibit the presentation of further results.

## 4 Conclusion

We describe a new super-resolution technique that uses a face mask to increase the resolution of non-planar objects despite changing viewpoints. Using this technique high-resolution images of objects in low-resolution video can be calculated during or after tracking. Experiments show that the proposed super-resolution algorithm increases resolution most significantly within the first 20 to 30 frames depending on the initial size of the face. Our combined geometric and appearance-based tracking algorithm is well suited for tracking faces of small size, using a 3D model mesh, which is a pre-requisite for the resolution enhancement.

## References


