Combining High-Resolution Images With Low-Quality Videos

Falk Schubert
EADS Innovation Works, Germany
falk.schubert@eads.net

Krystian Mikolajczyk
University of Surrey, UK
k.mikolajczyk@surrey.ac.uk

Abstract

Recently a lot of research has been directed towards the question: What can be done with "brute-force vision" using huge amounts of data? Image retrieval methods have been shown to succeed on collections of images with sizes over a million. Various applications such as object recognition, 3D geometrical arrangment of images showing the same scene or inferring missing image regions can benefit from large image databases. Motivated by this research we propose an alternative use of image information stored in large pools like the internet. Given an input video, we can utilize corresponding still images stored at much better quality to improve the overall quality of the video. A hybrid superresolution scheme is applied to smoothly incorporate the high-frequency components. On those areas where hallucination of details fails, a standard MAP-estimation of the high-resolution image is performed. The performance is demonstrated on real data examples.

1 Introduction

Much attention has been paid to the vast amount of information available through images in online databases like Flickr, Yahoo Image and Google Image Search in the past years. Methods for using image information contained in large image collections have been shown in work of [11, 19, 21]. Given the technology to handle such databases especially in terms of retrieval many new applications arise. *Hays et al.* [11] utilized the countless information of Flickr for scene-completion in a "brute-force vision" scheme. *Snavely et al.* [19] demonstrated an automatic and very robust sparse 3D reconstruction method, which formed the core of a 3D image browser. *Torralba et al.* [21] analysed what amount of data is actually needed for non-parametric "brute force vision" methods to work and what benefits arise using large databases for various vision tasks like classification, detection, etc.

This motivates the following application scenario: Assume we have a low-quality input video from a famous site for instance (e.g. Notredam Cathedral in Paris) that we wish to enhance. Using the above methodology one can retrieve a best matching image showing that same building in a similar view. We query the Flickr image database with certain keywords for a scene (e.g. "buckingham palace") and download about 500 images via the Flickr API. Usually it is easy to manually select a best fitting image. We also assume that these still images are usually available at much better quality than the input video. The goal is then to incorporate the details from the matched still image into the input sequence. Another application scenario with the same input constraints (low-quality

BMVC 2008 doi:10.5244/C.22.116

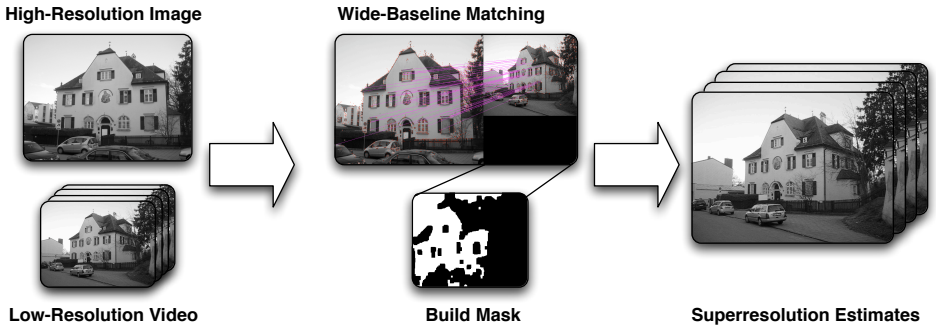

Figure 1. Outline of the proposed method: matching, masking, superresolution.

video and corresponding high-quality still image) could arise from special dual-mode cameras which capture a still image once in while (e.g. every second) while taking a video simultaneously as suggested in [12].

In this paper, we propose to use a combination of a reconstruction-based superresolution algorithm (MAP) with high-resolution prior information to merge both the given low-resolution image sequence and the high-resolution still image. Most likely the video and the still image were taken from different viewpoints. Applying a wide-baseline matching, one can align the still image to one of the low-resolution frames (e.g. the first frame). We assume static, planar scenes, hence in general scenes the alignment has to be performed in respect to the dominant plane (e.g. using homographies). This limitation is quite common given the inaccuracy (sub-pixel accuracy is required) and/or complexity (usually 3D dense stereo is employed) of general state-of-the-art registration methods like in [4]. We adapt histograms of oriented gradients [7] to compute a registration map that highlights areas showing the same scene. Since we compare high- and low-quality images which were possibly taken by different cameras, this masking must be robust to differences in illumination and image quality. MAP based superresolution using a generic prior on the unmasked parts and constraining the solution to additional high-resolution information provided by the still image on the masked is performed to merge the video and the high-quality still. In figure 1 the proposed method is outlined.

In section 2 a summary of state-of-the art superresolution methods is given. In section 3 the matching and mask building is described followed by the derivation of the superresolution estimation process in section 4. Results are shown in section 5.

2 Related Work

Details of images are mainly described by high-frequency parts of an image (e.g. edges). Using high-frequency information from still images to improve the quality of an input video or image has been proposed many times in the field of superresolution. Some methods employ generic collections of image patches of high-quality and match those to the input sequence [9, 5, 20]. Because of the generic type still images can only indirectly contribute to detail information. For instance similarly textured but different patches are augmented onto the low-resolution input (e.g. to enhance parts of a letter, patches drawn from images showing flowers might be used). Others use object-specific collections (e.g. only faces) [1] or use images taken in parallel to the input video for dynamically constructing such collections [12]. As a special case of the general spatio-temporal superres-

olution framework of [18] still images are combined with videos by temporal alignment to include information from both sources. *Bhat et. al* [4] proposed a general framework to transfer properties from still images to videos. A planarity assumption is not required for registration as they utilize a complex 3D dense stereo reconstruction. Among many other properties they considered to copy the greater amount of pixel information from the images to the video. However the fusion process resembles more a rerendering of the video frames mainly substituting pixels with ones from the still images rather than reconstructing to underlaying true scene.

In this work, we want to motivate the use of high-resolution information available in form of still images as prior constraints for reconstruction-based superresolution. Such a scenario can arise for instance when using the images of the internet as a knowledge source for instance. We demonstrate that it is possible to directly include detail information from the high-quality still image into the video. This circumvents the need of constructing patch databases and losing spatial correlation between patches (as done in learning-based superresolution approaches) and results in scene-specific image hallucination (e.g. to improve a video of a church, an image of that same church is used). A similar attention to globally constrain the use of a high-resolution image has been payed in superresolution of faces [6, 13].

3 Input Alignment

In order to merge pixel information from both inputs, the still image needs to be spatially aligned with the video. Assuming a static, planar scene (e.g. camera zooming and rotating around its center or far distance between scene and camera) or having a dominant planar scene to work on, we match robust features between the still image and a video frame for an approximate registration. The two inputs might have been recorded at different time points, hence only certain regions will overlap in both as real world scenes are usually subject to some form of dynamic change. Areas of moving objects or other structural changes in the scene can not be merged and need to be excluded. We automatically generate a binary mask specifying regions which are identical in both inputs and which are not. Using this mask a refined registration is performed on the unmasked areas. In order to process a whole video sequence, the registration with the still image is performed for all video frames. The binary mask is obtained using the first video frame and then aligned to the other video frames using the homographies from the last registration step.

3.1 Initial Registration

Given a high-resolution image $\vec{p}$ and a low-resolution frame $\vec{l}$, a mapping (e.g. homography, H) is sought so that a set of co-planar matches $\{(\vec{x}_i^h, \vec{x}_i^l)\}$ of feature points found in the high- and low-resolution image fulfills $H\vec{x}_i^h = \vec{x}_i^l, \forall i$. We employ SURF features [3] computed on each image and the standard homography estimation using RANSAC for outlier detection [10] to find such a mapping H. To cope with wide-baseline scenarios, the estimation is repeated multiple times (e.g. 3-5) using increasingly restrictive matching criteria and outlier thresholds.

3.2 Masking

In the next step areas of the aligned image pairs showing the same scene have to be robustly identified in an automatic fashion. The images might have been taken at different

time points, under varying viewing angle and by cameras with different quality. Hence besides structural changes in the scene (e.g. a person walked into the scene or a parked car has moved) there will also be differences in global illumination, local reflections, image quality and even some imperfect alignment in the initial registration step. A good similarity measure has to be robust against all these and yet be sensitive enough to distinguish between subtile differences in textured areas and structural changes in the scene. Besides a good initial geometric registration, a global photometric registration (contrast and brightness) is performed to reduce illumination differences. For simplicity no explicit temporal coherence is enforced in the current implementation. Therefore the robustness of the employed similarity measure is crucial for visually stable results (otherwise ghosting-artifacts may appear as shown in section 5).

Various similarity measures like sum-of-squared differences (SSD), mutual information (MI), normalized cross-correlation (NCC) and histograms of oriented gradients (HOG) can be employed. Following consideration have to be made for a good choice:
SSD is too sensitive to mis-alignment or illumination changes and not sensitive enough to similarly textured, but different areas – a single pixel difference does not give enough information.
MI is computed on a small region around each pixel. Generally mutual information is a very robust measure used for alignment of images coming from different sensors in medical imaging for instance. Hence it is insensitive to global differences in illumination. However it is difficult to estimate meaningful joint and marginal probabilities for small areas (10×10 pixels). Furthermore on homogeneous regions marginal entropies are very low resulting in an overall low mutual information value independent of whether these regions are similar or not.
NCC is also computed patchwise. It gives very low errors on well aligned structured areas (e.g. border of windows) but is still sensitive to local illumination changes (in figure 2 sky and homogeneously textured wall of house have almost same error as mis-aligned section of the roof because of slight illumination differences).

To measure the difference in structure and not in illumination, the measure should only consider high-frequency content. We adapt HOG features [7] computed on a dense grid to measure image similarity. The use of histograms achieves some robustness to noise via binning. They robustly represent image structures in a similar way to SIFT [15].

Some examples of high- and low-resolution input images and their various dissimilarity measures are shown in figure 2. The HOG measure gives the best mask for the two inputs. To generate a binary mask, the dissimilarity image needs to be binarized via thresholding. Erosion and dilation operations are applied to refine the mask. Finally the alignment of the high- and low-resolution image can be refined by applying a geometric intensity-based registration on the unmasked area. For this we employ the dual inverse compositional approach from [2], which serves as a compact framework for photometric registration (as applied in previous steps) and simultaneous geometric alignment. The mask building process is now repeated with slightly more sensitive parameter settings to obtain the final mask (see top of figure 2 for an example).

4 Superresolution

General multi-frame superresolution methods merge multiple low-resolution images (e.g. extracted from a video sequence) to form one high-resolution image. The resulting image

Figure 2. Top: Aligned sample input images (Left: first video frame, Right: still image). Bottom: different dissimilarity images between inputs (Left-Right: SSD, MI, NCC, HOG).

does not only contain a greater number of pixels (e.g. is upscaled in respect to the input), but it contains more information and detail than any of the input images alone. Various methods for this reconstruction problem have been developed in the past (see [6, 8, 16] for good overview). In the following a short summary on the mathematical background of general multi-frame superresolution is given.

4.1 Background

One way to solve the inverse problem of image reconstruction is based on iteratively minimizing a backprojection error: $err = \sum_i^N \rho(\vec{g}_i - \vec{l}_i)$, where ρ is a norm, $\vec{l}_i$ are the N observed low-resolution images and $\vec{g}_i$ are synthetically generated low-resolution images using the current estimate for the high-resolution image $\vec{h}$. An imaging model M contains all degradations that lead from the fine true high-resolution scene to the discretized low-quality version captured by the camera. Typically geometric transformations (caused by a moving camera), photometric changes, blur and downsampling are considered to constitute the capturing model. If the low-resolution images were extracted from a video sequence, only the geometric transformations are mostly different for each observed input image. By vectorizing the 2D image, one can formulate all those degradations as matrix-operations:

$$\vec{g}_i = \alpha \cdot D \cdot B \cdot T_i \cdot \vec{h} + \beta = M_i \cdot \vec{h} + \beta \tag{1}$$

α and β photometrically deform the image, D downsamples the image (by spatially averaging), B adds blur to the image and T_i geometrically transforms the image (usually the class of transformations is restricted to homographies for simplicity). The blur and downscale parameters have to be set manually. The transformation parameters (photometric and geometric) are computed by registering the frames $\vec{g}_i$ of the video. Since the interframe motion is usually small, we employ the same intensity-based method [2] as in the mask refinement step. Various formulations like maximum likelihood estimation (MLE) or maximum a posteriori formulation (MAP) [6] aim to maximize the probability $P(\vec{l}_i|\vec{h})$ that the high-resolution estimate generated the observed low-resolution images. Minimizing the negative log of this probability results in minimizing the following cost-function

for $\vec{h}$:

$$F_{standard} = \sum_i \|M_i \cdot \vec{h} - \vec{l}_i\|^2 + \text{constraints} = \|M \cdot \vec{h} - L\|^2 + \lambda \sum_{\vec{x}} \rho(\nabla(\vec{h},\vec{x}), \delta) \quad (2)$$

M and L are obtained by stacking all M_i and $\vec{l}_i$ onto eachother. Solving this ill-posed problem, requires sensitive constraints for obtaining good results. These are either referred to as regularization terms (in the MLE case) or prior contraints (in the MAP case). Typically the Huber-function $\rho(.)$ applied to the norm of the gradient of the superresolution estimate is used:

$$\rho(\nabla(\vec{h},\vec{x}), \delta) = \begin{cases} \left\|\nabla(\vec{h},\vec{x})\right\|^2 & \text{if } |\nabla(\vec{h},\vec{x})| < \delta \\ 2\delta|\nabla(\vec{h},\vec{x})| - \delta^2 \end{cases} \quad (3)$$

$\nabla(\vec{h},\vec{x})$ is the magnitude of the gradient of $\vec{h}$ at position $\vec{x}$. The Huber-function combines the smoothing Tikhonov regularization with the L2-norm ($\|.\|$) and the edge-preserving total variation regularizaton with the L1-norm ($|.|$) via a fix threshold δ.

4.2 Merging Inputs via Prior

The main goal of this part is to incorporate available prior high-resolution information into the superresolution estimation process in corresponding areas of the input. For this a binary mask was constructed in the previous step that indicates whether a pixel of the input has a corresponding high-resolution pixel in the high-resolution prior image or not. In the unmasked areas we want to use a generic edge preserving prior (e.g. Huber-prior) and in masked areas the solution shall be close to the corresponding high-resolution prior. Since the high-frequency are most responsible for sharp details perceivable by human eye (similar argument were used for instance in [9, 1]) and the low-resolution content from the input sequence shall be preserved to obtain a smooth result, the following additional gradient based prior is added to the cost function (2):

$$\|W(\nabla(\vec{h}) - \nabla(\vec{p}))\| \quad (4)$$

$\nabla(.)$ computes the gradients (vertical and horizontal) and W denotes the mask. Therefore the following final cost-function is subject to minimization:

$$F_{combined} = \|M \cdot \vec{h} - L\|^2 + \lambda \sum_{\vec{x}} \rho(\nabla(\vec{h},\vec{x}), \delta) + \gamma\|W(\nabla(\vec{h}) - \nabla(\vec{p}))\| \quad (5)$$

The first term is the reconstruction contraint from the MAP formulation (2). The second term refers to the generic image prior, the Huber-function (3) and the third term contains the high-resolution prior (4) using the high-quality input image $\vec{p}$. Using experimentally found values for the weights (λ, γ) for the priors and for the threshold δ, a smooth augmention of the additional high-resolution prior is achieved.

Because of the large numbers of equations, we employed the very fast limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [14] for optimization, but any non-linear optimizer like conjugated gradient descent could be used instead.

Usually only a few iterations are needed (30-100) until all residuals are minimized as shown in figure 3. Like [17] we initialize the optimizer with a MLE estimation (just the reconstruction constraint from (5) is optimized for a few iterations) which gives a much sharper starting point than the average image which is usually used [6]. We also allow for a registration refinement as suggested in [17], by iteratively solving equation (5) and registering the low-resolution input L to the current superresolution estimate $\vec{h}$. For registration we use the dual inverse compositional approach from [2]. Usually after

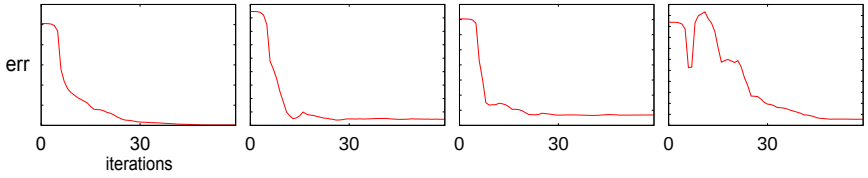

Figure 3. Individual residual errors along the iterations for a typical computation (Left-Right: total cost, reconstruction term, Huber-prior, high-resolution-prior).

Figure 4. First frame of input video and high-resolution image for "Notredam" sequence.

the first two iterations the registration refinement stops updating the warping parameters. Note that no explicit temporal coherence is enforced. However for the standard MAP superresolution temporal smoothness is achieved by combining multiple input images to one high-resolution image.

5 Results

The proposed algorithm was tested on various real sequences. The evaluation is performed in a qualitative measure comparing interpolation, standard MAP and the presented method[1]. The first video shows the "Notre Dame Cathedral". The camera is mostly rotating around its center justifying the use of homographies for the registration of the frames. We used the Flickr API to download 500 images labeled with "notre dame". The most suitable high-resolution image was manually selected from this set, although retrieval methods like [19, 11] could automate this step. The first frame of the video and the corresponding high-resolution image from Flickr are depicted in figure 4.

Because the video and the still image were taken from different viewpoints, the buildings on the left side of the cathedral are slightly moved. This part is highlighted by the mask so that no information of the high-resolution still image is included and only stan-

[1] see also http://personal.ee.surrey.ac.uk/Personal/F.Schubert/bmvc2008/ for results in full resolution

dard MAP superresolution is performed. However very fine details on the cathedral can only be recovered by the additional high-resolution image. The results of standard MAP superresolution on the whole image (10 frames were used to generate 1 high-resolution image) and the proposed combinations are shown in figures 5 and 6. The standard superresolution is capable of emphasizing edges and improving the quality of the image. Fine structure however can only be made visible by including additional knowledge of the details, for instance by querying the internet for a high-quality shot and including this information into the superresolution estimation process.

Figure 5. Left: Bicubic interpolation of low-resolution input, Middle: MAP superresolution using Huber-prior, Right: MAP superresolution with Huber-prior and high-resolution prior.

Figure 6. A zoom into images from figure 5 (marked by yellow rectangle in left image).

Figure 7. Left: high-resolution still image, Middle: low-resolution video frame, Right: MAP superresolution with Huber-prior and high-resolution prior.

Another example shows a poster of a car (see figure 7). The superresolution enhanced

images show fine details which could have never been recovered by standard superresolution alone. The still image is smoothly included into the reconstruction process (for example the lighting is also adjusted). In cases where the masking is incorrect, e.g. due to missing temporal coherence, ghost-edges appear as the MAP estimation tries to reconstruct the parts in the scene which are not really present (in figure 7 mask reaches over left border of the poster). Areas outside the mask (everything else than the poster) are processed with standard MAP superresolution and also show some quality improvement over the low-resolution input.

The third sequence shows a movie poster behind a window (see figure 8). It is straightforward to find high-quality still images of movie poster in the internet. Because of very strong reflections in the window, also parts of the poster are masked out (for instance parts of the headline text). However in regions where the high-resolution prior is applied a very strong quality improvement can be noticed. In the other regions slight enhancement by standard MAP superresolution is achieved.

Figure 8. Left: high-resolution still image, Middle: low-resolution video frame, Right: MAP superresolution with Huber-prior and high-resolution prior.

6 Conclusion

We motivate the use of high-resolution information in form of still images available from the internet as an additional prior for MAP-based superresolution on videos. The results demonstrate that it is possible to smoothly include the high-resolution images into the superresolution estimate. With robust registration the two inputs can be aligned and a masking procedure highlights parts that can be merged. Unmasked areas are enhanced using standard MAP superresolution. However accurate registration and masking of matching regions between the two inputs is crucial to avoid ghosting artifacts.

Possible applications might be inspired by systems like Phototourism [19], dual-mode cameras that simultaneously capture images and videos or by movie post-processing systems that enhance videos without the need to retake a scene. In general terms the proposed method gives an alternative way to combine prior knowledge from high-quality still images (e.g. images of popular sites from the web) with a low-quality video.

The results were evaluated by qualitative means as there does not exist a standarized measure or dataset to compare to. In the future we would like to investigate the performance in a more quantitative way using a larger range of videos.

Acknowledgment: This research was supported by UK EPSRC EP/F003420/1 grant.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *PAMI*, 24(9):1167–1183, 2002.

[2] A. Bartoli. Groupwise geometric and photometric direct image registration. In *PAMI*, accepted 2008.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[4] P. Bhat, C. L. Zitnick, N. Snavely, A. Agarwala, M. Agrawala, B. Curless, M. Cohen, and S. B. Kang. Using photographs to enhance videos of a static scene. In *EGSR*, 2007.

[5] C. M. Bishop, A. Blake, and B. Marthi. Super-resolution enhancement of video. In *AIS*, 2003.

[6] D. Capel. *Image-Mosaicing and Super-resolution*. Springer, ISBN: 1852337710, 2004.

[7] N. Dalal, N. Dalai, and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and challenges in super-resolution. *IJIST*, 14(2):47–57, 2004.

[9] W.T. Freeman, W.T. Freeman, T.R. Jones, and E.C. Pasztor. Example-based super-resolution. *CGA*, 22(2):56–65, 2002.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004.

[11] J. Hays and A. A. Efros. Scene completion using millions of photographs. *SIGGRAPH*, 2007.

[12] D. Kong, M. Han, W. Xu, H. Tao, and Y. H. Gong. Video super-resolution with scene-specific priors. In *BMVC*, 2006.

[13] C. Liu, H.-Y.g Shum, and C.-S. Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *CVPR*, 2001.

[14] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(3):503–528, 1989.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] T.Q. Pham. *Spatiotonal Adaptivity in Super-Resolution of Under-sampled Image Sequences*. PhD thesis, Delft University of Technology, 2006.

[17] L. C. Pickup, S. J. Roberts, and A. Zisserman. Optimizing and learning for super-resolution. In *BMVC*, 2006.

[18] E. Shechtman, E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *PAMI*, 27(4):531–545, 2005.

[19] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH*, 2006.

[20] J. Sun, J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum. Image hallucination with primal sketch priors. In *CVPR*, 2003.

[21] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical Report MIT-CSAIL-TR-2007-024, CS and AI Lab, MIT, 2007.