Super-resolved Digests of Humans in Video

Dong Seon Cheng, Marco Cristani, Vittorio Murino

Super-resolved Digests of Humans in Video

Dong Seon Cheng, Marco Cristani, and Vittorio Murino

Dipartimento di Informatica, University of Verona
Strada le Grazie 15, 37134 Verona, Italy
{dongseon.cheng,marco.cristani,vittorio.murino}@univr.it

Abstract. This paper describes a hierarchical approach towards the extraction of highly informative summarized information of humans from video sequences. Objects of interest, such as facial features, are detected through transformation-invariant clustering of the frames, iteratively from bigger to smaller regions, and then expressed with an information-rich representation obtained by super-resolution. To guarantee the fundamental constraints under which the super-resolution process is well-behaved, we propose a Bayesian framework that integrates the uncertainties in the registration of the frames. The ultimate product of the overall process is a strip of images that describe at high resolution the dynamics of the video, switching between alternative local descriptions in response to visual changes.

Key words: Super-resolution, video analysis

1 Introduction

Video summarization considers the problem of generating a concise and expressive summary of a video sequence by extracting and abstracting the most relevant features in the scene. In another context, super-resolution (SR) techniques aim at recovering a high resolution (HR) image starting from a set of low resolution (LR) frames, which are considered generated by the HR one [1,2,3,4,5]. This process can be thought as inverting the generative equation in which the LR frames are generated from the HR one when correctly warped, subsampled and blurred by the Point Spread Function (PSF) of the camera device. These two techniques can be joined together into a novel application that distills few HR surrogates from several similar (in a generative sense) noisy LR elements, and this is the aim of the paper.

In the classical SR framework, the common hypothesis is that LR images must be registered through sub-pixel displacements; this can be accomplished by knowing or inferring all the transformations (rigid or not), in order to correctly invert them. Moreover, once fixed a desired level of magnification, the number of LR images should be sufficient for recovering high frequency information. These fundamental constraints are quite hard and penalizing, making the super-resolution image estimation possible only under supervised and limited conditions, strongly reducing its applicability in wider contexts.

In this paper, we show how it is possible to overcome the fundamental SR constraints in video sequences, using a hierarchical clustering procedure. Roughly speaking, the identification of objects with persistent appearances might not be enough to guarantee correct sub-pixel registrations on the *whole* area covered by the objects. In fact, it is a common feature of most interesting objects (e.g., human beings) that they present moving parts or regions that abruptly change appearance. It makes sense, then, to cut those regions that present high variability and treat them hierarchically as a new sequence of images to be analyzed locally. In this way, the complete hierarchy made of alternating clusters and regions maintains a consistent level of good local registrations between frames that allows super-resolution to be effective in integrating the remaining uncertainties. At the end of the process, the set of all super-resolved visual patches are combined together to produce a highly resolved visual summarization of the entire sequence.

To be more precise, the idea consists first in fixing all the possible transformations to which the LR images are subjected, using a discrete approximation formed by invertible rigid translation and rotations. Over this space it is possible to perform transformation invariant component analysis, as presented in [6]. In such clustering framework, each cluster represents a particular object, through a mean that subsumes its main visual appearance, a subspace representation that spans its appearance variability (*i.e.*, due to local variations) and the covariance matrix that mirrors the uncertainty due to discrepancy between the image formation process and the model.

This discrepancy ultimately comes from three sources:

1. differences in pixel values among registered LR frames due to sub-pixel different displacements of the captured scene;
2. wrong modelling of the transformations which the object of interest is subjected to;
3. genuine changes in appearance that cannot be modelled through motion transformations.

The first kind of uncertainty can be used to immerse the LR space in a denser one and hence estimate sub-pixel shifts. The second noise is minimized only when the whole image is subject to discrete transformations, and conversely it increases when the LR images are subjected to local transformations. In this case, it helps to consider partitioning the LR space into smaller sub-spaces, over which the clustering process is performed again. This recursive hierarchical process stops when the patch examined is spanned by a fragment of the object of interest only via local rigid transformations. Then, at this point, a local registration can be performed.

This local clustering-registration step permits us to recover the fundamental hypotheses required to place our data in a super-resolution framework. Therefore, the actual super-resolution task is carried out in Bayesian fashion via a modified version of the EM procedure, using the grouped local patches of each sub-cluster and actively embedding an accuracy measure based on the covariance matrix obtained at the end of the hierarchical process, in order to build

higher resolved versions for each patch, taking differently into account the value of each pixel involved in the process. The idea is that the more present a value in the normalized versions of the grouped LR patches, the more probably this pixel holds an important role in the HR image reconstruction step.

At the end, the resulting process represents a novel kind of video summarization in which all the generated high resolution images can be registered against a common coordinate system by applying the several inverse transformations that affect local patches.

The rest of the paper is organized as follows. In Sec. (2) the state of the art of the SR methods is presented. In Sec. (3), the bayesian approach to super resolution is reported. The clustering approach and the modified approach to super resolution is detailed in Sec. (4). In Sec. (5), the method is tested and results are shown. Finally, in Sec. (6), conclusions are drawn and future perspectives are envisaged.

2 State of the art

The literature concerning the classical SR (in the sense that all the above constraints are met, also named here signal-based SR or reconstruction based [7]) is large and multifaceted, although two subgroups can be devised: in the first one, the alignment of the LR images is separated from the fusion step, that estimates the PSF parameters [2]; in the second group, all the parameters are jointly estimated [3,4,8]. Here the HR estimation is either performed by a maximum-likelihood (ML) approach, or in a Maximum A-Posteriori (MAP) fashion, regularizing the ill-conditioning of the ML framework using some prior [4,8]. In [7], the limits of the reconstruction-based SR algorithms is reported, in dependence of registration error, denoising accuracy, high frequency details in the HR image. Moreover, it gives also the sufficient number of LR images to estimate an HR image with magnification factor M. For a fractional M, the sufficient number is $4M^2$; for a integer M, the sufficient number is M^2. In [9], an innovative approach is presented, that is based on an irregular displacement of pictorial details of the LR images, called Penrose tiling. This approach outperforms several existing reconstruction-based algorithms for regular pixel arrays, and has performances which are not described by the constraints expressed in [7]. In [10], the handled registration model is fully projective, incorporating also a photometric model to handle brightness changes present in the frames of a temporal sequence. Additionally, the algorithm learns parameters for the regularizer high-resolution image prior.

An additional SR group of approaches can be devised, i.e. the learning-based approaches [11,12]. Learning-based SR algorithms are techniques that do not process images at the signal level; instead, they use case-dependent priors to infer missing details in low resolution images (LRIs). The seminal papers of this class of approaches are the methods proposed in [12] and the "Hallucination" algorithm [11]. In practice, the basic idea is that they study similarity (usually, per-patch) among the input LR image and LR training images, collecting to-

gether their correspondent high resolution versions. Once the similarities have been computed, the HR images related to the similar training LR images are employed to statistically infer the high frequency details of the super resolved version of the input image. The inference is usually cast as a MAP estimation. In general, pros of the learning-based SR approaches are that they work on fewer LRIs but can still achieve a higher magnification factor than traditional algorithms can. Most of them can even work on a single image. Cons are that, usually, the magnification factor is usually fixed and the performances depend on the matching with the training low resolution samples. Because of their advantages, learning-based SR algorithms have become popular. In [13] the limits of the learning-based super-resolution approaches for natural images are estimated, exploiting the statistics of the natural images, and ignoring the effects of the image noise. Such limits are modelled as an upper bound of the magnification factor, such that the expected risk (RMSE between the HR input image and the HR ground truth) is below a relatively large threshold. Moreover, they provide a formula that gives the sufficient number of HRIs to be sampled in order to ensure the accuracy of the estimate.

3 The Bayesian super-resolution formulation

Let the observed data consists in K low-resolution images $\mathbf{x}_k$, $k = 1, \ldots, K$, stored as one-dimensional column vectors obtained by raster-scanning. We assume these images have the same size $L_x \times L_y$, and hence the same number of pixels $L_n = L_x L_y$.

The classical super-resolution formulation describes how this observed data is generated from the unknown high-resolution image $\mathbf{z}$, with $H_n = H_x H_y$ pixels, where $H_x = qL_x$ and $H_y = qL_y$, depending on the linear magnification factor $q > 1$. In the following, we will assume q is set to an integer value.

The likelihood of observing each image $\mathbf{x}_k$ is modeled by the following Gaussian probability density:

$$p(\mathbf{x}_k|\mathbf{z}, \mathbf{s}_k, \theta_k, \gamma) \sim \mathcal{N}(\mathbf{x}_k; \mathbf{W}_k\mathbf{z}, \boldsymbol{\Upsilon}), \tag{1}$$

where $\mathbf{s}_k$ is the global sub-pixel shift, θ_k is the orientation displacement and γ regulates the width of the PSF.

The image formation process is neatly rendered by the projection matrix $\mathbf{W}_k$, which takes the high-resolution image $\mathbf{z}$ and warps it according to $\mathbf{s}_k$ and θ_k, blurring and down-sampling according to the PSF and the magnification factor. Finally, $\boldsymbol{\Upsilon} = \epsilon^2\mathbf{I}$ represents the residual post-transformation noise variance.

In particular, $\mathbf{W}_k$ is an $L_n \times H_n$ matrix with elements $w_{i,j}^{(k)}$ given by

$$w_{i,j}^{(k)} \propto \exp\left\{\frac{||\mathbf{v}_j - \mathbf{u}_i^{(k)}||^2}{\gamma^2}\right\}, \tag{2}$$

where the vector $\mathbf{v}_j$ represents the spatial position of the j-th HR pixel.

The vector $\mathbf{u}_i^{(k)}$ is the center of the PSF and is located according to the following transformation:

$$\mathbf{u}_i^{(k)} = \mathbf{R}_k(\mathbf{u}_i - \overline{\mathbf{u}}) + \overline{\mathbf{u}} + \mathbf{s}_k, \tag{3}$$

where the position $\mathbf{u}_i$ of the i-th LR pixel is rotated around the center $\overline{\mathbf{u}}$ of the image space by the standard rotation matrix $\mathbf{R}_k$, expressing the orientation displacement θ_k, and then shifted by $\mathbf{s}_k$.

To complete the Bayesian characterization of the super-resolution model, the following Gaussian regularization prior $p(\mathbf{x})$ over the HR image constrains the values of nearby pixels:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{C}), \tag{4}$$

where the covariance matrix $\mathbf{C}$ has the following $c_{i,j}$ elements:

$$c_{i,j} = \frac{1}{A} \exp\left\{ -\frac{\| \mathbf{v}_i - \mathbf{v}_j \|^2}{r^2} \right\}. \tag{5}$$

This prior enforces some smoothness on the high-resolution image by linking every i-th and j-th pixel values according to their relative locations in the image space and the parameters A and r, which represent the strength and range of the correlation, respectively.

Assuming flat priors on the model parameters $\{\mathbf{s}_k, \theta_k\}_{k=1}^K$ and γ, the posterior over the high-resolution image is proportional to the product of the prior and the joint likelihood terms:

$$p(\mathbf{z}|\{\mathbf{x}_k, \mathbf{s}_k, \theta_k\}_{k=1}^K, \gamma) \propto p(\mathbf{z}) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{z}, \mathbf{s}_k, \theta_k, \gamma). \tag{6}$$

After some matrix manipulations (see [14]), observing that we have a conjugate prior, it is possible to rewrite this posterior in the following Gaussian form:

$$p(\mathbf{z}|\{\mathbf{x}_k, \mathbf{s}_k, \theta_k\}_{k=1}^K, \gamma) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{7}$$

where

$$\boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \epsilon^{-2} \sum_{k=1}^{K} \mathbf{W}_k^T \mathbf{W}_k \right)^{-1} \tag{8}$$

$$\boldsymbol{\mu} = \epsilon^{-2} \boldsymbol{\Sigma} \left(\sum_{k=1}^{K} \mathbf{W}_k^T \mathbf{x}_k \right) \tag{9}$$

This posterior is basically a prior-compensated pseudo inverse matrix. The covariance in Eq. (8) encodes the uncertainty over each HR pixel: this uncertainty is mainly driven by the smallest value between the prior covariance and the covariance of the likelihood term, weighted with noise variance ϵ^2.

Given this Bayesian formulation of the super-resolution problem, learning is achieved through the EM algorithm[15], treating the low-resolution images $\mathbf{x}_k$ as the observed variables and the high-resolution image $\mathbf{z}$ as an hidden variable.

The EM algorithm iterates between two steps: in the E-step, given the current estimates of the model parameters $\{\mathbf{s}_k, \theta_k\}_{k=1}^K$ and γ, it computes new optimal estimates for $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$; in the M-step, it updates the estimates for the model parameters.

4 Super-resolved digests

In this section, we describe firstly our proposed procedure to recover the fundamental conditions under which the image super-resolution problem is well-behaved starting from a video sequence, and secondly how to synthesize a highly informative summarization of it.

Analysis of a video. Let $\mathbf{x}_k$, $k = 1, \ldots, K$, be the K frames of a given video sequence, treated as a set of images. Our first goal is to isolate the interesting visual objects, which we call *objects of interest* (OOIs): in the case of videos showing people, our primary OOIs are the persons figures. We assume, in general, that an OOI cannot be described by a single rigid appearance, but can be explained adequately by breaking it down hierarchically into parts, themselves broken down into sub-parts if needed, which we call secondary and tertiary OOIs. We informally describe each OOI as a set of alternative visual appearances and dynamics.

Our proposed procedure iteratively employs two operations: *DetectOOI* and *SplitOOI*, respectively to isolate OOIs and to break down an OOI into parts. For the sake of simplicity, in the following we will describe these two operations on the data $\mathbf{x}_t$, delaying to the next section the pre-processing necessary on real videos.

In *DetectOOI*, we use mixture of transformation-invariant component analyzers (MTCA)[6] to perform transformation-invariant clustering of the frames $\mathbf{x}_k$ (each frame having L_n pixels) and learn a subspace representation within each cluster. From a generative point of view, given C clusters with parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\Phi}_c, \boldsymbol{\Lambda}_c\}_{c=1}^C$, where $\boldsymbol{\mu}_c$ is the mean cluster image, $\boldsymbol{\Phi}_c$ is a diagonal covariance, and $\boldsymbol{\Lambda}_c$ is an $L_n \times F$ factor loading matrix, we first generate a latent image $\mathbf{z}_k$ with the following distribution:

$$p(\mathbf{z}_k|\mathbf{y}_k, c_k) = \mathcal{N}(\mathbf{z}_k; \boldsymbol{\mu}_c + \boldsymbol{\Lambda}_c\mathbf{y}_k, \boldsymbol{\Phi}_c), \tag{10}$$

where $\mathbf{y}_k$ is a F-dimensional Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ random variable representing the subspace coefficients.

The observed image $\mathbf{x}_k$ is then obtained by applying a transformation $\mathbf{T} \in \mathcal{T}$ on the latent image $\mathbf{z}_k$ and adding independent Gaussian noise $\boldsymbol{\Psi}$:

$$p(\mathbf{x}_k|\mathbf{z}_k, \mathbf{T}_k) = \mathcal{N}(\mathbf{x}_k; \mathbf{T}_k\mathbf{z}, \boldsymbol{\Psi}), \tag{11}$$

where the set of transformations $\mathcal{T}$ the model is invariant to must be specified a priori.

Model learning in MTCA is performed with an EM algorithm [6] starting from the joint distribution over all variables:

$$p(\{\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k, \mathbf{T}_k, c_k\}_{k=1}^K) = \prod_{k=1}^K p(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k, \mathbf{T}_k, c_k) = \tag{12}$$

$$= \prod_{k=1}^K p(\mathbf{x}_k|\mathbf{z}_k, \mathbf{T}_k) p(\mathbf{z}_k|\mathbf{y}_k, c_k) p(\mathbf{y}_k) P(\mathbf{T}_k) P(c_k) = \tag{13}$$

$$= \prod_{k=1}^K \mathcal{N}(\mathbf{x}_k; \mathbf{T}_k\mathbf{z}, \boldsymbol{\Psi}) \mathcal{N}(\mathbf{z}_k; \boldsymbol{\mu}_c + \boldsymbol{\Lambda}_c\mathbf{y}_k, \boldsymbol{\Phi}_c) \mathcal{N}(\mathbf{y}_k; \mathbf{0}, \mathbf{I}) \pi_{\mathbf{T}_k} \pi_{c_k}. \tag{14}$$

In *DetectOOI*, after MTCA is performed, interpreting each cluster as an OOI, we can split the observed K frames based on the MAP values for c_k and invert the MAP transformations $\mathbf{T}_k$, to effectively obtain C new data sets with image frames registered on the visual objects whose mean appearances are $\boldsymbol{\mu}_c$. Each OOI image also comes with a subspace representation $\boldsymbol{\Lambda}_c\mathbf{y}_k$, whose coefficients $\mathbf{y}_k$ allow us to study the dynamics of structural changes in the appearance of the visual object.

The diagonal covariance matrix $\boldsymbol{\Phi}_c$ contains the residual variance not explained by the subspace representation. We latch onto this "imperfections" in the *SplitOOI* operation to decide if the OOI is well characterized as it stands or needs to be broken down. These imperfections exist in the first place only if the number F of hidden factors is limited by design to a small fixed number. It is the same design decision as choosing to keep only F principal components, accounting for only a fraction of the total variance.

In *SplitOOI*, we isolate image regions where the residual variance is higher than a given threshold, which can be decided a priori or computed from the distribution of the actual values. This operation potentially determines multiple sites where the OOI appearance dynamics cannot be explained by the learned subspace representation. As an example, it might occur that a *face* OOI contains *eyes* and *mouth* regions that are badly modeled.

We determine compact and connected regions by using simple small-valued morphological operators (closing, dilation) and proceed to cut the image frames of the given OOI based on the bounding box of each detected part. This gives rise to new data sets, each representing a secondary OOI, with reduced size and number of images.

At this point, *DetectOOI* and *SplitOOI* are performed again to detect, if they exist, tertiary OOIs. After each *SplitOOI* the procedure might be halted when some conditions are met, for example if there are no sizable residual variance regions or there is no reduction in size, meaning MTCA has failed to identify and characterize consistent clusters. See Fig. (1) for a graphical representation of a typical detect and split analysis process.

Improved super-resolution. Note how this OOI parameterization is ideal for our intended use of image super-resolution: the frames have been aligned pixel-wise on a common visual object, only the frames actually containing such

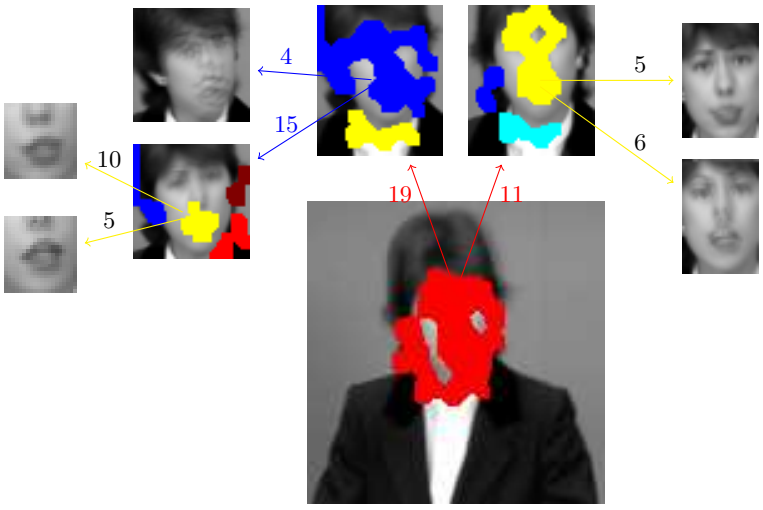

Fig. 1. Graphical representation of a typical detect and split process for the analysis of a video sequence. The colored blobs indicate OOIs that have been tagged for splitting. Here are shown only the most interesting splits, with the images showing average appearances and the numbers on the arrows indicating the partitioning of the frames.

visual object have been selected, subspace analysis has detected the principal structural changes, and the residual variance tell us where super-resolution may fail.

For the sake of simplicity, let $\mathbf{x}_l$, $l = 1, \ldots, L$, be the L transformation corrected images, of size $O_x \times O_y$, belonging to cluster c. We propose to modify Eq. (1) to take into account the residual cluster variance $\boldsymbol{\Phi}_c$ in the following way:

$$p(\mathbf{x}_l|\mathbf{z}, \mathbf{s}_l, \theta_l, \gamma) \sim \mathcal{N}(\mathbf{x}_l; \mathbf{W}_l\mathbf{z}, \boldsymbol{\Phi}_c + \boldsymbol{\Upsilon}), \tag{15}$$

where the new covariance term $\boldsymbol{\Phi}_c + \boldsymbol{\Upsilon}$ in the LR image formation process expresses sensor noise contributions and contributions from the OOI detection process, specifically indicating sites that may need more "slack", or, in other words, should be trusted less to derive a correct sub-pixel registration.

After some matrix manipulations, and denoting $\boldsymbol{\Delta} = \boldsymbol{\Phi}_c + \boldsymbol{\Upsilon}$, we can update Eqs. (7)(8) and (9) in the following way:

$$p(\mathbf{z}|\{\mathbf{x}_l, \mathbf{s}_l, \theta_l\}_{l=1}^{L}, \gamma) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{16}$$

where

$$\boldsymbol{\Sigma}' = \left(\mathbf{C}^{-1} + \sum_{l=1}^{L} \mathbf{W}_l^T \boldsymbol{\Delta}^{-1} \mathbf{W}_l\right)^{-1} \tag{17}$$

$$\boldsymbol{\mu}' = \boldsymbol{\Sigma}' \left(\sum_{l=1}^{L} \mathbf{W}_l^T \boldsymbol{\Delta}^{-1} \mathbf{x}_l\right) \tag{18}$$

Fig. 2. Six frames from the "Claire" video sequence.

Moreover, from the analysis process we can also derive a rough measure of similarity between the frames in a given cluster, useful to further select the most useful images for super-resolution. Let $\mathbf{Y} = [\mathbf{y}_1 \mathbf{y}_2 \dots \mathbf{y}_L]$ be the $F \times L$ matrix obtained by assembling the subspace coefficients of all L images in a given cluster c, then the square $L \times L$ matrix $\mathbf{Y}^T\mathbf{Y}$ represents a similarity matrix defined on the inner product in the subspace $\boldsymbol{\Lambda}_c$.

Reasoning on $\mathbf{Y}^T\mathbf{Y}$ allows us to compare frames within a cluster, to identify blocks of similar frames, to identify the most representative frames, or, given a particular frame, to select the closest matchings. Note that, if we are within a very compact cluster, such conclusions might be inappropriate, overblowing the significance of tiny details, that may very well be useful for super-resolution.

Super-resolved digests. Finally, all the gathered information after the analysis phase can be used to synthesize a highly informative summary of the given video sequence. Firstly, we take each cluster and perform image super-resolution, with the intent of representing all the frames of such clusters with single images at a higher resolution, summarizing their information in a more compact way than the cluster parameters (mean and variance).

Secondly, we build the timeline of switchings between the different clusters at the different hierarchical levels, so as to know which super-resolved images must be combined to rebuild a complete-frame reconstruction starting from the broken down representation we built in the analysis phase. See Fig. (4) for an example of super-resolved digest.

5 Experimental Session

In Fig. (2), we show several frames from the "Claire" test video sequence, in its short version of about 30 frames sized 176×144. To speed up the analysis, we perform some simple pre-processing on the video frames to eliminate most static background, on the assumption that in this case the background is not a candidate object of interest. In particular, we first collect pixel statistics and isolate the overall bounding box of the high variance sites as foreground. Moreover, the first *DetectOOI* stage is performed on modified frames where the background has been substituted with random values drawn from a uniform distribution, we call this procedure "taping", with the intent of leading the learning process to ignore those regions. The original frames are reintroduced before the following *SplitOOI* stage.

In Fig. (3), we show a detailed view of the analysis that was also reported in Fig. (1). This analysis is mostly focused on the detection of the different poses

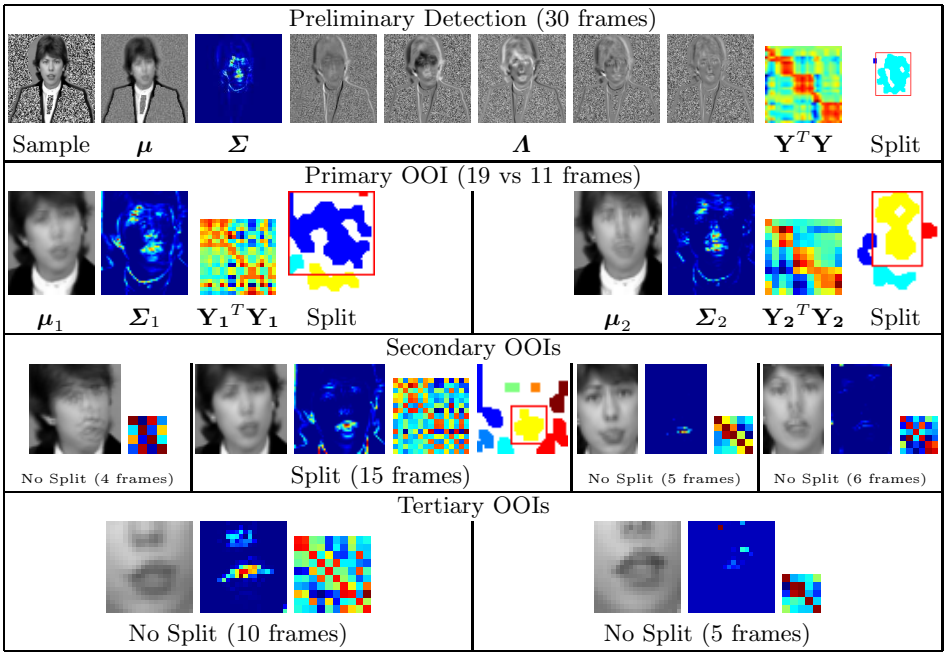

Fig. 3. Detailed view of the analysis performed on the "Claire" video. In the first row, from left to right, a sample "taped" image (see text) followed by the single cluster mean, variance, hidden factors, similarity matrix and split targets. The following rows present the mean, variance and similarity matrices of the hierarchical decomposition of the video. Split target maps are given for those OOI that have been split, with the results in the following row.

and expressions of the given person. To avoid falling into a bad local maximum of the log-likelihood, each *DetectOOI* operation was performed ten times and the try achieving the best log-likelihood was retained for further splits or detections. Also, for this experiment we considered only translational motions to ease the computations, but we reserve to model rotations, scaling and other warpings in future improvements.

By looking at the similarity matrix in the preliminary phase (second to last diagram in the first row of Fig. (3)), it looks like there are four main blocks of similar frames plus a small transitioning phase before the last block. This fact becomes evident in the final super-resolved digest in Fig. (4), where the transition cluster is actually pretty garbage. In this digest, no frames were eliminated prior to the super-resolution process, leading to high-resolution images that are slightly imperfect. To avoid this problem, we can examine the similarity matrix, either to automatically choose the biggest group of similar frames, or to score frames starting from a given reference frame and eliminating those that score too poorly (see Fig. (5)). Overall, we decided to use all the available frames, given that the process is generally well behaved.

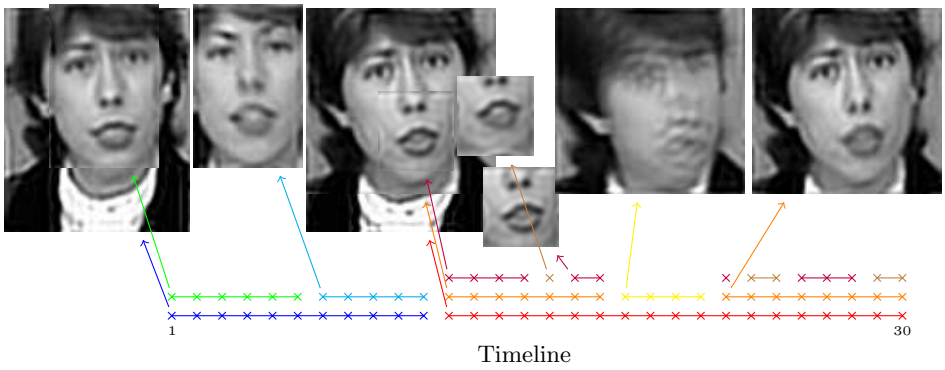

Fig. 4. Super-resolved digest of the "Claire" video. At the bottom, the reconstructed timeline of "events" that trigger the cluster switchings. Linked to the main events are the super-resolved frame reconstructions on the top.

Fig. 5. Super-resolution performed without frame elimination (on the left) and eliminating frames given a reference frame (on the right).

6 Conclusions

In this paper, we have proposed a novel framework to provide video analysis through transformation-invariant clustering and super-resolution. In this framework, we define several uncertainties intrinsic in video sequences portraying complex objects of interests and determine a scheme to identify and separate them in such a way as to make super-resolution effective and meaningful. This analysis leads immediately to the composition of reconstructed super-resolved digests of the elements in the video, similar to the process of video summarization. This is particularly well-suited for the portrayal of human figures. The preliminary experiments carried out have shown results that seem promising for further research. In fact, we envisage future advancements, like integrating the several independent subspaces into a global characterization or using them to synthesize high-resolution motions, and practical applications, like video synthesis and image stabilization.

References

1. Schultz, R., Stevenson, R.: A Bayesian approach to image expansion for improved definition. IEEE Trans. on Image Processing **3**(3) (1994) 233–242
2. Cheeseman, P., Kanefsky, B., Kraft, R., Stutz, J., Hanson, R.: Super-resolved surface reconstruction from multiple images. In Heidbreder, G.R., ed.: Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, Dordrecht, the Netherlands (1996) 293–308
3. Hardie, R., Barnard, K., Armstrong, E.: Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. IEEE Trans. on Image Processing **6** (1997) 1621–1633
4. Tipping, M., Bishop, C.: Bayesian image super-resolution. In: Neural Information Processing Systems, Vancouver, Canada (2002)
5. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Comput. Graph. Appl. **22**(2) (2002) 56–65
6. Kannan, A., Jojic, N., Frey, B.: Fast transformation-invariant component analysis. Int. J. of Computer Vision **77**(1-3) (2008) 87–101
7. Lin, Z., Shum, H.: Fundamental limits of reconstruction-based superresolution algorithms under local translation. IEEE Trans. on Pattern Analysis and Machine Intelligence **26**(1) (2004) 83–97
8. Pickup, L.C., Capel, D.P., Roberts, S.J., Zisserman, A.: Bayesian image super-resolution, continued. In: Neural Information Processing Systems, Vancouver, Canada (2006)
9. Ben-Ezra, M., Zhouchen, L., Wilburn, B.: Penrose pixels super-resolution in the detector layout domain. In: Proc. IEEE Int. Conf. on Computer Vision, Rio de Janeiro, Brazil (2007)
10. Pickup, L.C., Roberts, S.J., Zisserman, A.: Optimizing and learning for super-resolution. In: Proc. of the British Machine Vision Conference, Edinburgh, UK (2006) 439–448
11. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. on Pattern Analysis and Machine Intelligence **24**(9) (2002) 1167–1183
12. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. Int. J. of Computer Vision **40**(1) (2000) 25–47
13. Lin, Z., He, J., Tang, X., Tang, C.: Limits of learning-based superresolution algorithms. In: Proc. IEEE Int. Conf. on Computer Vision, Rio de Janeiro, Brazil (2007)
14. Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. J. of Machine Learning Research **1** (2001) 211–244
15. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B **39** (1977) 1–38