Automatic Generation of Subject-Based Image Transitions

Edoardo Ardizzone, Roberto Gallea, Marco La Cascia, and Marco Morana

Università degli Studi di Palermo
{ardizzon,robertogallea,lacascia,marcomorana}@unipa.it

Abstract. This paper presents a novel approach for the automatic generation of image slideshows. Counter to standard cross-fading, the idea is to operate the image transitions keeping the subject focused in the intermediate frames by automatically identifying him/her and preserving face and facial features alignment. This is done by using a novel Active Shape Model and time-series Image Registration. The final result is an aesthetically appealing slideshow which emphasizes the subject. The results have been evaluated with a users' response survey. The outcomes show that the proposed slideshow concept is widely preferred by final users w.r.t. standard image transitions.

Keywords: Face processing, image morphing, image registration.

1 Introduction

In recent years, the diffusion of digital image acquisition devices (e.g., mobile phones, compact digital cameras, smartphones) encouraged people to take more and more pictures. However, the more pictures are stored the more users need automatic tools to manage them. In particular, personal photo libraries show peculiar characteristics as compared to generic image collection. Personal photos are usually captured on the occasion of real-life events (e.g., birthdays, weddings, trips), so that it is common the presence of people in most of the images. Moreover a relatively small number of different individuals (i.e., the family) is usually shown across the whole library and this allows to achieve reliable results with automatic approaches.

Many methods for photo collection management were proposed focusing on users' request for accessing a subset of stored data according to some particular picture properties.

In this work we address the scenario where the user wants to manage its own photo collection according to *who* is depicted in each photo. Starting from a collection of personal images, we propose a tool for the automatic slideshow of a sequence of pictures depicting the same individual. Firstly the whole collection is searched for faces, while face identities are assigned with an automatic approach [1]. Then each face is processed for automatically finding the position of some facial feature points. Finally, image sequences that contain the same person are animated by applying a morphing approach.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 237–246, 2011.

The problem of the automatic morphing of pairs of digital images has been investigated since the early '90s. More recentlty, several techniques have been focused on face morphing. In [2] and [3] two systems are proposed for the automatic replacement of faces in photographs using Active Shape Models (ASMs). Both of them use standard ASMs on simple images, i.e., high-quality frontal faces, thus it seems that such approaches are not robust enough to be used with faces captured "in the wild". This is a strong limitation since even if a number of features detection techniques are available in literature, their performances signitificaly drop when such methods are applied on real-life images.

Some commercial systems, such as *Fantamorph* [4], allows users to create face animation. Although its latest version provides support for face and facial features detection, user intervention is often needed to aid the detection process and no face identification is available.

The paper is arranged as follows: a description of the system will be given in Sect. 2. The Sect. 2.1 will give an overview of the face processing techniques we developed, while the image alignment methods are described in Sect. 2.2. Experimental results will be shown and discussed in Sect. 3. Conclusions will follow in Sect. 4.

To better evaluate the results of the proposed work, sample videos of produced slideshows are available at `http://www.dinfo.unipa.it/~cvip/pps/`.

2 Methods

The whole system realizing the image animation can be subdivided into two main blocks (see Fig. 1): face processing and image alignment modules. The first is responsible of detecting and identifying the subject, the second operates the spatial transformation across the transitions.

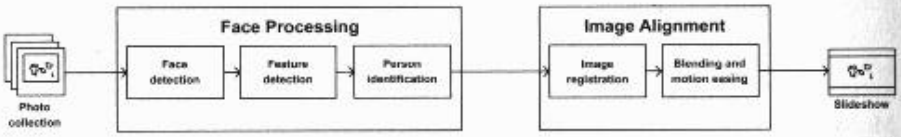


Fig. 1. Block diagram of the proposed system

2.1 Face Processing

Face processing is considered nowadays as one of the most important application of image analysis and understanding.

Face detection, i.e., determine if and where there is a face in the image, is usually the first step of face processing techniques. A number of challenges are associated with face detection due to several factors. For example, face appearance may heavily change according to the relative camera-face pose, to the presence of facial hair, or on the occurrence of lighting variations. All these factors are further stressed in real-life photo collections.

Many face detection techniques [5] have been proposed and even if face detection is not a solved problem, some methods have reached a certain level of maturity. In this work we used the state of the art approach to face detection, i.e., the framework proposed by Viola and Jones [6], due to its efficiency and classification rate.

When making the animation of two subsequent pictures, as much local information as possible is needed to preserve the appearance of face regions. Thus, the corners of the bounding box obtained from the face detection step are not sufficient and local facial features need to be detected.

Early approaches for facial feature detection focused on template matching to detect eyes and mouth [7], but these features are not suitable for noisy images such as real-life photos. More recent models, i.e., ASM, AFM, AAM, offer more robustness and reliability working on local *feature points* position. Active Shape Model (ASM) [8] extends Active Contour Model [9] using a flexible statistical model to find feature points position consistently with the training set. Active Appearance Model (AAM) [10] combines shapes with gray-level face appearance.

Due to their time efficiency, ASMs are one of the most used approach both for face detection and real time face tracking. However the efficiency of this approach is heavily dependent on the training data used to build the model, that in most cases consists of face databases (e.g., FERET, Color FERET, PIE, Yale, AR, ORL, BioID) acquired under constrained conditions. Such collections of faces looks very different from those acquired in everyday life, thus in the considered scenario this approach is unsatisfactory. For this reason we used personal collections both for training and testing. The training set is a collection of faces detected in a private personal collection, while the whole system has been tested on a publicly available dataset [11] enabling future comparison.

Once faces have been detected, we focused on using ASMs to find a predefined set of fiducial points in face images. Considering a training set of 400 size-normalized faces (200x200 pixels) detected in a private photo collection, we developed three models, shown in Fig. 2, using 45, 30 and 23 landmarks respectively. The first model, Fig. 2(a), is composed of five shapes, i.e., right eye, left eye, mouth, nose-eyebrows and face profile. This model is frequently used in ASMs-based works, however we performed some tests on real-life pictures noticing that in most cases, due to lighting variations, face profile and internal shapes (mouth and eyes) are misplaced while the nose-eyebrow contour appears as the most robust to pose changes. For this reason the first model has been refined as shown in Fig. 2(b). In this case we considered a single shape by removing the tip of the nose and linking mouth and eyes contours. However we discovered that the tip of the nose is fundamental for the right positioning of surrounding features, moreover the areas around the eyes are frequently noisy due to hard intensity changes caused by occlusions (e.g., glasses). Consequently, the contours of the eyes are not reliable enough.

Taking into account such considerations, our final model (Fig. 2(c)) consists of a single shape that follows the top contour of the eyebrows and the bottom contour of nose and mouth. A comparison of the feature detected with the three models is shown in Fig. 2.

Fig. 2. The three developed models. In leftmost column red points represent the first and last landmark for each shape. Other columns depict examples of features detection. Each row shows the shapes obtained by using the first, second and third face model respectively.

The 23 detected points are used as input of the morphing algorithm. These points, in conjunction with information about *who* is in the picture (i.e., the face identity) allow to perform the automatic animation of the photo sequence.

Organizing the collection based on *who* is in the photo generally requires much work from the user that, in the worst case, has to manually annotate all the photos in the collection. Here we used a data association framework for people re-identification [1] that takes advantage of an important constraint: a person can not be present two times in the same photo and if a face is associated to an identity, the remaining faces in the same photo must be associated to other identities. The problem is modeled as the search for probable associations between faces detected in subsequent photos using face and clothing descriptions. In particular, a two level architecture is adopted: at the first level, associations are computed within meaningful temporal windows (events); at the second level, the resulting clusters are re-processed to find associations across events.

The output of the re-identification process is a set of identities associated to each face detected in the collection.

2.2 Image Alignment and Photo Transitions

Given the correspondences between two consecutives images, the next task to be performed is the smooth transition bringing the first images onto the second,

keeping the main object (i.e., the face of the considered person) aligned. Such problem can be regarded as a time-series registration task, or more specifically, a morphing problem. Some issues need to be considered for this purpose: choosing the registration function for the alignment, the easing function for the transition and the motion function for the feature points.

Registration Function. During the whole transition, in order to mantain spatial coherence, the feature points need to be kept in alignment. Such problem is defined image registration. It can be considered as the geometric function to apply to the image I (input image), to bring it in correspondence with the image R (reference image). Many approaches exist for such purpose, in this work we adopted two strategies. The first one is based on region-wise affine registration, the latter leverages onto thin-plate spline transforms.

In the case of *region-wise multi-affine transformation*, the main idea is to subdivide the images in sub-regions, each one defined by three feature points, resulting corresponding triangles are aligned using affine transformations. Points triangulation in I and R is obtained using the classic Delaunay triangulation.

Since the vertices of the obtained triangles tessellation and their respective displacements are known, for each triangle is possible to recover the mapping function for all the points belonging to it. Each 2d affine transformation requires six parameters, so writing down the transformation (both x and y coordinates) for three points, produces a system of six equations, sufficient for its complete determination, as stated in (1).

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a\ b\ c \\ d\ e\ f \\ 0\ 0\ 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} = \begin{bmatrix} ax_0 + by_0 + c \\ dx_0 + ey_0 + f \\ 1 \end{bmatrix} \Rightarrow \begin{cases} x_1 = ax_{0,1} + by_{0,1} + c \\ y_1 = dx_{0,1} + ey_{0,1} + f \\ x_2 = ax_{0,2} + by_{0,2} + c \\ y_2 = dx_{0,2} + ey_{0,2} + f \\ x_3 = ax_{0,3} + by_{0,3} + c \\ y_3 = dx_{0,3} + ey_{0,3} + f \end{cases} \quad (1)$$

Each transformation is recovered from a triangle pair correspondence, and the composition of all the transformations allows the full reconstruction of the image. In addition, to avoid crisp edge across triangles edges, the *fuzzy kernel regression* approach [12] was used for transformation smoothing.

The second approach is based on *thin-plate spline transformations* (TPS) [13]. The TPS is a parametric interpolation function which is defined by $D(K+3)$ parameters, where D is the number of spatial dimensions of the datasets and K is the number of the given landmark points where the displacement values are known. The function is a composition of an affine part, defined by 3 parameters, and K radial basis functions, defined by an equal number of parameters. Its analytic form is defined as:

$$g(p) = ax + by + d + \sum_{i=1}^{k} \rho\left(\|\mathbf{p} - \mathbf{c}_i\|^2\right) w_i; \qquad \mathbf{p} = \begin{bmatrix} x_p \\ y_p \end{bmatrix}, \mathbf{c}_i = \begin{bmatrix} x_{c_i} \\ y_{c_i} \end{bmatrix} \quad (2)$$

where $\mathbf{p}$ is the input point, $\mathbf{c}_i$ are the landmark points and the radial basis function $\rho(r)$ is given by:

$$\rho(r) = \frac{1}{2} r^2 \log r \tag{3}$$

All of the TPS parameters are computed solving a linear system defined by a closed-form minimization of the bending energy functional. Such functional is given by:

$$E_{tps} = \sum_{i=1}^{k} \|y_i - g(\mathbf{p}_i)\| + \lambda \int\int \left[\left(\frac{\partial^2 g}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial xy} \right)^2 + \left(\frac{\partial^2 g}{\partial y^2} \right)^2 \right] dxdy. \tag{4}$$

The functional is composed by two terms: the data term and the regularization term. The former minimizes the difference between known and recovered displacements at landmark points, the latter minimizes the bending energy of the recovered function, i.e., maximises its smoothness and it is weighted by the parameter λ. As mentioned before, for this expression a closed-form analytical solution exists, from which is possible to recover all of the required spline function parameters. The main characteristic of this function is that it exhibits minimum curvature properties.

Notwithstanding the used deformation function, for the morphing animation purposes, at each time step both the two images A and B involved in the transition need to be registered to an intermediate image defined by the current frame considered, whose feature points lie onto the path connecting their corresponding feature points (Fig. 3).

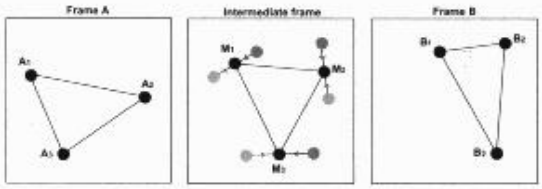


Fig. 3. When transitioning from frame A to frame B, both feature points of A and B, need to move to intermediate feature points to reconstruct intermediate image M

Easing Function. The visual transition between the two images is realized through a blending of the two images registered to the intermediate images. Such blending is produced as the weighted sum of the intensity level of the images, where the weighting factor is defined by a blending paramter b. Varying the parameter in function of the time during the animation, produces an easing visual effect. The easiest variation criterion is to adjust it linearly with the time variation. However, linear easing is equivalent to no easing at all, this is quite unaesthetic since the variation occurs instantly, resulting in a crisp bad-looking

mechanical effect. For this reason, for determining $b(t)$, several non-linear easing functions based on Penner's formula [14] have been implemented.

Basically, three types of easing are used, ease-in (slow start with instantaneous stop), ease-out (instantaneous start and slow stop) and ease-in-out (slow start and stop), using four function shapes, quadratic, cubic, sinusoidal and exponential. In Fig.4 are reported the 12 resulting different easing functions.

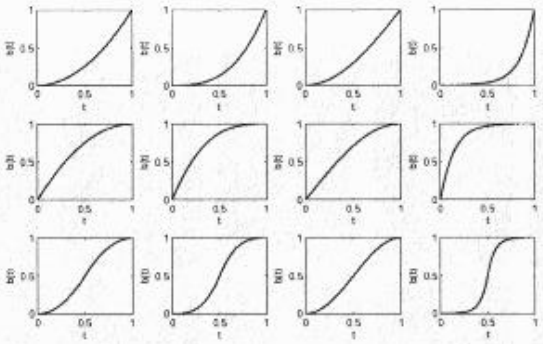


Fig. 4. Easing functions used for the determination of the blending parameter $b(t)$ and the position parameter $p(t)$. From left to right: quadratic, cubic, sinusoidal and exponential easing. From top to bottom: ease-in, ease-out, ease-in-out functions.

Motion Function. As for blending, an easing factor for the path followed by the feature points during the transition is introduced too. Such parameter $p(t)$, determines the equation of motion of the points as a combination of feature positions in image A and B. The same functions in Fig.4 are used for computing $p(t)$.

3 Results and Discussion

For the proposed approach, a quantitative evaluation is neither possible nor required, since the purpose of the project is to produce aesthetically appealing animations for subject-based slideshows. Thus, the evaluation was just qualitative. However, statistics of people opinions about the images were collected. In particular, we created a benchmark of videos and conducted a user study to compare our method to standard slideshows. Then, we present an analysis of users' responses, finding an agreement on the evaluation of the results.

3.1 Video Benchmarks

We created a benchmark of videos based on the images in the Gallagher Collection Person Dataset [11]. The videos were realized using 24 frames per second

and 5 image transitions with a running time of 1 second each. For each test case two videos were generated, one using our slideshow approach, and one using standard cross-fading. In Fig.5 some frames of a video are shown. Top and bottom rows depict the videos generated with the proposed and the standard approach respectively. In particular frame 1, 7, 13, 19 and 24 are illustrated.

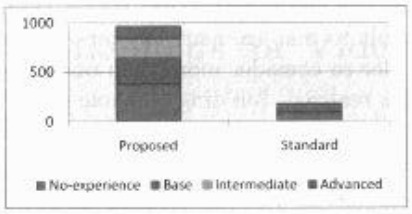
To better evaluate the results of the proposed work, sample videos of produced slideshows are available at `http://www.dinfo.unipa.it/~cvip/pps/`.

Fig. 5. Frame 1, 7, 13, 19 and 24 of an example of benchmark video for the proposed slideshow (top row), and the standard cross-fading slideshow (bottom row)

3.2 Users' Evaluation

In order to accomplish user evaluation, a user's reponse analysis system was developed, relying onto the web-based survey system provided by the *RetargetMe* framework [15]. Once the system is set up, the survey is taken by 113 people with different background in image analysis spanning from *no experience* to *advanced*. For each survey 10 test cases are submitted. For each test case two videos are presented and the person is requested to choose which one is the preferred. Alternatively only one video can be presented and the person is requested to give an absolute evaluation with a vote between 1 and 5.

In Fig. 6 the outcome of the survey is reported: averagely about the 80% of the people preferred the proposed version of the transition, while this percentage increase for people very familiar with image analysis and processing concepts. Fig.7 illustrates statistics of the votes given from an absolute point of view. The boxplot diagram shows that the mean vote for the proposed slideshow (≈ 4.33) is higher than the vote for the standard one (≈ 3.32). The motivation given when preferring the standard slideshow is generally related to the presence of



Background	People	Proposed	Standard
No-exp.	45	373 (82.89%)	77 (17.11%)
Base	33	269 (81.52%)	61 (18.48%)
Intermediate	23	203 (88.26%)	27 (11.74%)
Advanced	12	114 (95.00%)	6 (05.00%)
Total	113	959 (84.87%)	171 (15.13%)

Fig. 6. Statistics of users preferences about the proposed slideshow and the cross-fading slideshow. Stacked bars are used to express the different backgrounds of people who took the survey.

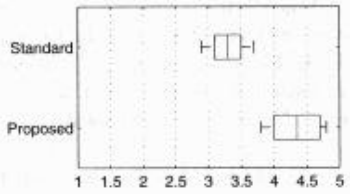


Fig. 7. Boxplot diagrams of absolute evaluation, expressing the votes assigned to the videos created with the proposed slideshow (bottom row), and the standard cross-fading effect (top row)

deformations in non-subject regions of the images, which are needed to preserve alignment of subject regions. However, this should not be considered a system fault, since this issue is inherent to the system design itself.

4 Conclusion and Future Works

A novel method for fully automatic subject-based slideshow generation was presented. The method is aimed to realize a cross-fade image transition which keep focused the person representing the main subject of the pictures. This is achieved by means of aligning and morphing the subject face (along facial features) while realizing the transition. This allows to keep the attention of the user onto the subject, attaining a pleasant and aesthetically attractive visual result. Given a set of images, the system recognizes and locates the subject, then performs a time-varying registration (i.e., a morphing) to mantain it aligned during the animation. Such alignment is realized using a deformation function based on triangles mesh with locally affine transformations or thin plate spline surfaces. In addition easing functions are used to give a more natural and pleasant aesthetic look to the transitions and object movement.

The system was evaluated using a web-based framework submitted to heterogeneous audience, which judged the videos according to the visual appealing, both through comparisons and absolute assessment. The survey response was

that the proposed slideshows are more appealing, providing a nice effect that can be implemented as a feature in personal photo management software.

The project can be further extended in order to consider more than one person, in addition other warping effects can be realized, building a whole set of transition plugins.

References

1. Lo Presti, L., Morana, M., La Cascia, M.: A data association algorithm for people re-identification in photo sequences. In: 2010 IEEE International Symposium on Multimedia (ISM), pp. 318–323 (2010)
2. Min, F., Lu, T., Zhang, Y.: Automatic face replacement in photographs based on active shape models. In: Asia-Pacific Conference on Computational Intelligence and Industrial Applications, PACIIA 2009, vol. 1, pp. 170–173 (2009)
3. Terada, T., Fukui, T., Igarashi, T., Nakao, K., Kashimoto, A.: Yen-Wei Chen. Automatic facial image manipulation system and facial texture analysis. In: Fifth International Conference on Natural Computation, ICNC 2009, vol. 6, pp. 8–12 (2009)
4. Fantamorph. Abrosoft (2010), `http://www.fantamorph.com`
5. Yang, M.-H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002)
6. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)
7. Yuille, A.L., Cohen, D.S., Hallinan, P.W.: Feature extraction from faces using deformable templates. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings CVPR 1989, pp. 104–109 (June 1989)
8. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models—their training and application. Comput. Vis. Image Underst. 61(1), 38–59 (1995)
9. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision V1(4), 321–331 (1988)
10. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, p. 484. Springer, Heidelberg (1998)
11. Gallagher, A., Chen, T.: Clothing cosegmentation for recognizing people. In: Proc. CVPR (2008)
12. Gallea, R., Ardizzone, E., Gambino, O., Pirrone, R.: Multi-modal image registration using fuzzy kernel regression. In: ICIP 2009, pp. 193–196 (2009)
13. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 567–585 (1989)
14. Penner, R.: Programming Macromedia Flash MX. McGraw-Hill, New York (2002)
15. Rubinstein, M., Gutierrez, D., Sorkine, O., Shamir, A.: A comparative study of image retargeting. ACM Transactions on Graphics (SIGGRAPH) 29(5) (2010)