Video Segmentation for Markerless Motion Capture in Unconstrained Environments

Martin Côté[1], Pierre Payeur[1], and Gilles Comeau[2]

[1] School of Information Technology and Engineering
[2] Department of Music
University of Ottawa
Ottawa, Ontario, Canada, K1N 6N5
{mcote,ppayeur}@site.uottawa.ca, gcomeau@uottawa.ca

Abstract. Segmentation is a first and important step in video-based motion capture applications. A lack of constraints can make this process daunting and difficult to achieve. We propose a technique that makes use of an improved JSEG procedure in the context of markerless motion capture for performance evaluation of human beings in unconstrained environments. In the proposed algorithm a non-parametric clustering of image data is performed in order to produce homogenous colour-texture regions. The clusters are modified using soft – classifications and allow the J-Value segmentation to deal with smooth colour and lighting transitions. The regions are adapted using an original merging and video stack tracking algorithm.

1 Introduction

Image segmentation is often considered one of the most important low level vision processes. It has recently been extended from colour images to video sequences with field applications in video encoding and video database indexing [1, 2]. The concept of representing video regions in terms of objects has also been introduced. An analysis of these objects can provide more insight as to the content and the semantics of a video. In this particular case, objects representing individuals could be evaluated to extract information regarding their activities. The provision of quantitative measurements for human performances using a passive vision based system has a strong appeal for activities in the field of music and sports where performance measurements are based on human perception and experience. This type of application is often referred to as Motion Capture.

Recently there has been significant advancement in the field of computer vision techniques. However, none have yet addressed the complex problem faced here without having to impose unreasonable constraints upon musicians or athletes and their environments. Many motion capture techniques using passive sensors still rely on contrasting backgrounds or on assumptions on the motion and complexity of the scene. These impositions yield an environment that is foreign to a performer, potentially compromising the integrity of his actions, leading him to behave differently than he would in a more comfortable environment. The limitations of such techniques may also obfuscate key performance markers through the application of arbitrary data

G. Bebis et al. (Eds.): ISVC 2007, Part II, LNCS 4842, pp. 791–800, 2007.

representations or manipulations. We introduce the concept of unconstrained environments, where a performer and his environment are faced with a minimum of assumptions and requirements allowing him to perform uninhibited.

Two categories of segmentation techniques are explored in the application of a Motion Capture system. The first category deals with frames in a sequential manner. This field has been explored thoroughly and has too many varying approaches to list within the scope of this paper. Some of the more popular techniques can be categorized as contour-based, background modeling and region space approaches. In the case of contour-based approaches, techniques are often driven by image gradients in order to produce a delineation of important image components. One of the founding techniques, called active contours, was introduced by Kass *et al.* [3]. The contours are formed using an energy minimization procedure designed in such a way that its local minima are achieved when the contour corresponds to the boundary of an object. The technique was modified for video objects by Sun *et al.* [1] using a projective tracking algorithm but is not well suited for large non-rigid movements. In the case of background modeling techniques, one successful algorithm was introduced by Stauffer and Grimson [4]. They proposed the use of a mixture of Gaussian probability models to capture individual pixel behaviours and separate active foreground objects from low-motion background objects. Despite various improvements to this technique [5], in the case of performance evaluation, assumptions on the presence of motion cannot always be made, making the distinction between foreground and background objects complex. Finally, region-based approaches perform an analysis of the data space in order to produce a simplified grouped representation of the data. The union of these regions makes the process of segmentation and tracking much simpler. In the case of watershed algorithms [6, 7, 8], regions are formed by identifying local minima within a frame's gradient image. More adaptive techniques achieve segmentation using an adaptation of the k-means algorithm [9]. The criterion used for the creation of regions can yield different results depending on the nature of the images processed. The second category of segmentation techniques deals with frames in sets of blocks, called video stacks, and has received an increasing amount of attention. In DeMenthon [10], video stack segmentation uses a modified Mean-Shift approach which is computationally intensive, requiring a hierarchical implementation.

The hybrid approach proposed within this paper incorporates a video stack analysis with a sequential frame tracking of segmented video objects. It avoids the high computational and memory cost of volume based analysis by separating the video stream into frame windows. A combination of clustering and spatio-temporal segmentation techniques is performed on the video window in order to extract pervasive homogenous regions. The algorithm builds upon the JSEG approach introduced in [11] and extended by Wang *et al.* [12].

2 General Approach

The proposed technique is categorized as a region-based motion capture segmentation algorithm and uses colour-texture information to produce homogenous regions within a set of frames that are then tracked throughout the sequence. The technique is based on Deng and Manjunath's JSEG implementation [11] with key improvements making

it more appropriate to the context of a performer evaluation considered here. The algorithm is structured as a set of five key processes: clustering, soft-classification, J-value segmentation, merging and tracking. While many of these processes have been addressed in the original JSEG algorithm, this work proposes several improvements and introduces algorithms which have shown to be more efficient within the harsh environments that we tested in.

2.1 Non-parametric Clustering of Images

As a precursor to the actual segmentation the video stacks must first undergo a clustering process. Originally proposed by Deng *et al.* [11] was a k-means based approach which assumes that the colours present within a scene follow Gaussian-like statistics. This hypothesis cannot always be guaranteed for complex scenes. Wang *et al.* [12] also reached this conclusion and modified the approach to use a non-parametric clustering technique called the Fast Adaptive Mean-Shift (FAMS). The FAMS algorithm introduced by Georgescu *et al.* [13] builds upon the original Mean-Shift technique proposed by Comaniciu *et al.* [14]. It is used within our approach to cluster colour distributions within a video stack without applying constraints to these distributions. Only the basic concepts of the Mean-Shift property and the FAMS algorithm are conveyed here.

Given n data points such as $x_i \in R^d$, $i = 1,\ldots,n$ associated with a bandwidth $h_i > 0$, the multivariate kernel density estimator at location x is defined as:

$$\hat{f}_{h,k}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \tag{1}$$

Where $k(x)$ is a function defining the kernel profile and $c_{k,d}$ is a normalization constant. If the derivative of the kernel profile $k(x)$ exists, a density gradient estimator can be obtained from the gradient of the density estimator yielding the following:

$$\nabla \hat{f}_{h,k}(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (x_i - x) k'\left(\left\| \frac{x - x_i}{h} \right\|^2 \right) = \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} k'\left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} x_i k'\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} k'\left(\left\| \frac{x - x_i}{h} \right\|^2 \right)} - x \right] \tag{2}$$

The last term in equation (2) is called the Mean-Shift. This term, by definition, points in the direction of the maximum increase in the density surrounding a point x. By applying the mean shift property iteratively, we converge on the mode of a given point. By associating the mode of each distribution to the data points converging to it, a nonparametric clustering of the data space is obtained.

In [13] several other improvements were brought to the clustering technique. These include the use of adaptive bandwidth sizes and an optimization technique called Locality-Sensitive Hashing (LSH) that aims to speed up the clustering process. This speed up requires a lengthy pre-processing in order to obtain optimal parameters that would yield the best computation time and reduced error. In the implementation

done here, the adaptive bandwidths were omitted and optimization parameters were manually selected. These omissions did not overly affect the clustering process but did allow for a much quicker processing. The end result is an algorithm that achieves a better colour clustering in light of smooth colour gradients.

2.2 Creation of Soft-Classification Maps

In the list of improvements to the original JSEG algorithm, Wang *et al.* [12] introduced the concept of soft-classification maps. These maps represent a measured membership value that a pixel has to its assigned cluster. These values allow the JSEG algorithm to soften the colour-texture edges between two similar cluster distributions. The classification maps can be created for every pixel using Bayesian probabilities. The cluster distributions in this case are represented using Gaussian statistics in order to compute the corresponding memberships.

In this paper, we have opted to compute our classification maps differently. The use of Gaussian statistics to describe the clusters undermines the idea behind the use of the non-parametric FAMS algorithm. Instead, the clusters are represented using 3D normalized histograms of Lu*v* pixel intensities. The values of the histogram bins, when projected back into the image, represent the non-parametric probability, $P(I_k \mid w_i)$, that pixel I_k belongs to the class w_i. This process is called a histogram back-projection [2] and allows for the creation of soft-classification maps without the need to assume particular distributions on the clusters.

2.3 J-Value Segmentation

JSEG is a novel segmentation technique that attempts to produce regions out of pixel labelled images. In this case, the labels are generated by the FAMS process described earlier and represent the distribution assigned to each pixel. The first step in the segmentation is to compute a homogeneity measure of every pixel based on its neighbours. This measurement depicts the local variation in colour classification surrounding a pixel. This value, called the J-value, is presented here following the same notation as adopted by Deng *et al.* [11] and adapted in order to take into consideration the previously defined soft-classification maps. First the mean position of clusters is defined as:

$$m = \frac{1}{N}\sum_{z \in Z} z \tag{3}$$

Where m is the mean, Z the set of all N data points within a local region around a pixel and $z = (x, y), \quad z \in Z$. Assuming that there are a total of C colour clusters, we can define the mean position of a particular cluster i as:

$$m_i = \frac{\sum_{z \in Z} z \cdot \omega_{z,i}}{\sum_{z \in Z} \omega_{z,i}} \qquad i = 1,..., C \tag{4}$$

Here $\omega_{z,i}$ is the membership value taken from the soft-classification maps defined in the previous section. The introduction of this term is a modification proposed by [12] to the original JSEG technique and allows the membership values to influence the

mean position of a particular cluster. Finally, the total spatial variance of clusters is defined as:

$$S_T = \sum_{z \in Z} \|z - m\|^2 \tag{5}$$

Similarly, the sum of all cluster variances is given as:

$$S_W = \sum_{i=1}^{C} S_i = \sum_{i=1}^{C} \sum_{z \in Z} (\omega_{z,i} \cdot \|z - m_i\|^2) \tag{6}$$

The term $\omega_{z,i}$ also makes an appearance within the above variance computation and allows the same membership values to play a role within the computation of S_W. The J-value of the local region is obtained based on these variances:

$$J = (S_T - S_W) / S_W \tag{7}$$

The original paper [11] provides examples on how a particular local cluster distribution would affect the outcome of the J-value. For a local region where clusters are distributed approximately uniformly, the J-value will remain relatively small. Inversely, should the local region consist of segregated clusters; the J-value will increase. The result of an image wide J-value computation is a gradient image corresponding to homogeneous colour-texture edges.

The set of points which define the local region on which the J-value is computed is described by a circularly symmetric kernel mask. This mask is applied to every pixel in an image. The kernel size depends on the scale at which J-values are computed. At a larger scale smoother texture edges are detected while at a smaller scale, hard edges are detected. The process is iterative; once regions are determined at a large scale, the regions undergo another JSEG process in order to split them based on a smaller kernel size. Regions are created using a seed growing algorithm that amalgamates nearby pixels having a low J-value.

The JSEG algorithm also allows for video segmentation by way of seed tracking. The tracking algorithm presented by Deng *et al.* [11] requires that all video frames be segmented at once and depends on small motion between frames. This is not practical for very large or lengthy videos; a solution to this is presented within section 2.5. JSEG also defines the term J_t in order to measure spatio-temporal homogeneity. This term, computed similarly as its J counterpart, helps to indicate which pixels should be used when determining seed overlap. Only pixels with a good spatio-temporal homogeneity are considered.

2.4 Joint-Criteria Region Merging

Both the JSEG and modified JSEG approaches suffer from over-segmentation. Its original authors proposed a simple merging algorithm that iteratively attempted to merge regions having the closest corresponding histograms. Over-segmentation being a classical problem, it has been explored extensively within other contexts [6, 7, 8]. We adopted an algorithm that uses a joint space merging criterion introduced by Hernandez *et al.* [6]. This technique not only relies on colour information but also on the edges between two candidates. As such, it prevents the accidental merging of regions with similar colour attributes having a strong edge in between them.

The first step in performing the merge operation is to formulate a Region Adjacency Graph (RAG) [15]. Region labels are represented by graph nodes while their similarities with adjacent regions are represented by edges. Region merges are done iteratively and invoke an update of the RAG. The similarity criterion used is based on both colour homogeneity and edge integrity. Colour homogeneity is defined as a weighted Euclidian distance between the colour means of two adjacent regions. The weight is computed based on region sizes and will favour the merging of smaller regions. The edge integrity criterion is based on the ratio of strong edge pixels and regular edge pixels found along the boundary of two adjacent regions. In order to compute this ratio, a gradient image is first created using Wang's [8] morphological method. A threshold is found based on the median value of the gradient image. Any pixels found to have a value higher than the threshold are considered strong boundary pixels. The edge criterion will increase in the case where two regions are separated by a prominent edge.

To produce a single similarity criterion, both homogeneity and edge integrity criteria must be evaluated. Since their scales are not known, Hernadez *et al.* [6] suggest using a rank based procedure where the final similarity is given by:

$$W = \alpha R^H + (1-\alpha) R^\varepsilon \tag{8}$$

Here R^H and R^ε are the respective ranks of the criteria given above for the same two adjacent regions. α is a weight parameter used to impart importance on either of the former criteria.

2.5 Region Tracking

The tracking algorithm developed for this framework combines the strengths of sequential and video stack segmentation to create a hybrid strategy to track regions. The resulting technique is described in the following sections.

2.5.1 Intra-video Stack Tracking

The original JSEG algorithm allows for block segmentation with the use of a seed tracking procedure. This tracking however requires that the video be segmented in its entirety by considering all the frames at once in order to be successful and is often not feasible due to memory constraints. We propose to first separate the segmented video in a series of video stacks. The size of the stacks can be manipulated by an operator and depends on the available memory and computing power. The tracking and region determination applied within a video stack is done using the technique proposed by Deng *et al.* [11] and described in section 2.3. The tracking done in between stacks is described in the next section.

2.5.2 Inter-video Stack Tracking

The inter-video stack tracking algorithm proposed is strongly based on region overlaps between two consecutive video frames. This means that the motions exhibited by the objects in the video must be captured with an appropriate frame rate in order to allow regions to have an overlap between frames. The tracking correspondence indicator used within this work stems from the research produced by Withers *et al.* [16]. In their work, the authors have tried to identify region correspondences between

frames regardless of splitting, merging and non-uniform changes. This tracking methodology lends itself well to the segmentation technique presented here.

The criterion used to find a correspondence between regions of two subsequent frames depends highly on distance and pixel overlap. In this case, pixel overlap is defined as the number of pixels one region has in common with another between two frames. Withers *et al.* [16] define the overlap-ratio, $R_{i,j}(t)$, as the correspondence measure between region i and j, it is given by the following equation:

$$R_{i,j}(t) = \frac{V_{i,j}(t)}{D_{i,j}(t)} \tag{9}$$

Here the terms $D_{i,j}(t)$ and $V_{i,j}(t)$ are distance and overlap ratios for the intersection of regions i and j. These ratios are defined as the fraction between the distance and overlap of regions i and j respectively and those of the region intersection exhibiting the smallest value. Regions that may have undergone a splitting or merging will still have a very large overlap-ratio with their ancestors. By applying a threshold to the overlap-ratio, eq. (9), final correspondence can be achieved.

3 Experimental Results

This section provides a comparison between the various additions proposed in this paper and the steps found in the original JSEG algorithm. It also presents sample results of the final segmentation that can be achieved. Due to the nature of the algorithm and of the improvements, it is difficult to define quantitative evaluation metrics that would apply to such segmentation methods. However, the improvements are easily discernable by comparing the clusters of pixels. All sequences were captured at a resolution of 320x240 and 30 fps and depict piano players performing in complex environments.

In Figure 1 a clustering comparison between the original k-means and FAMS is shown. The major disadvantage with the k-means algorithm is that it requires extensive parameter tweaking in order to obtain a good clustering. In the first sequence the k-means algorithm does not produce nearly as many clusters as FAMS, many of the image details are lost in the over clustering. In the second sequence, FAMS is better able to distinguish colours from various image components. The piano is described using fewer clusters. In the case of the pianist a distinction between the left and right arms as well as the torso can be made. These improvements are in part due to FAMS's ability to account for colour gradients, thus allowing cluster centers to differ from their actual mean. The figure also looks at the quality of the results when parameters are manually selected. FAMS requires more time to complete than its k-means counterpart, but gives better results. If an optimization on the parameters selection is performed, the overall computation time is significantly increased for a negligible difference in results.

Figure 2 shows the effect of the soft-classification of cluster on the J-value computations. Larger J-values are represented by brighter pixels. The figure depicts

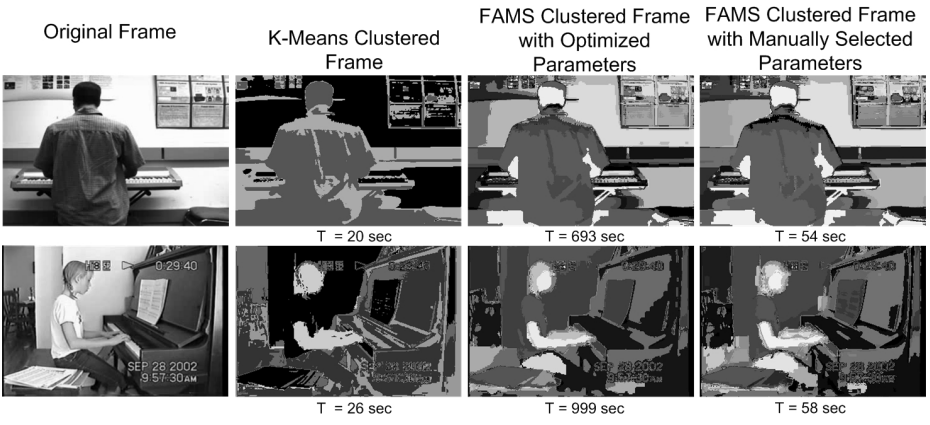

Fig. 1. K-Means and FAMS Comparison

Fig. 2. Impact of Soft-Classification

the results of two soft-classifications computed using the smallest kernel size with Gaussian distributions [12] and the proposed histogram back-projection. The soft-classification of clusters results in a softening of the non-homogeneous colour-texture edges. The use of Gaussian distributions leads to an over attenuation of J-Values, while the more flexible histogram back-projection yields J-Values that do not remove key image details. The attenuation of values ultimately results in fewer seeds being generated for nearby regions having similar colour-texture properties.

The results in Figure 3 clearly demonstrate a reduction in superfluous regions caused by colour gradients and lighting effects. The shaded regions are the ones selected by a human operator and are relevant to the motion capture process. In the first video where more clusters were found using FAMS, a better outline of the musician is achieved. In the second video where the number of clusters was approximately the same, the regions form better contours and identify semantic image components clearly. In particular the pianist's torso, arms and legs can be identified more easily. The change from Gaussian to histogram based soft-classification allowed regions to better keep their distinctive shapes and conform to the semantic video content.

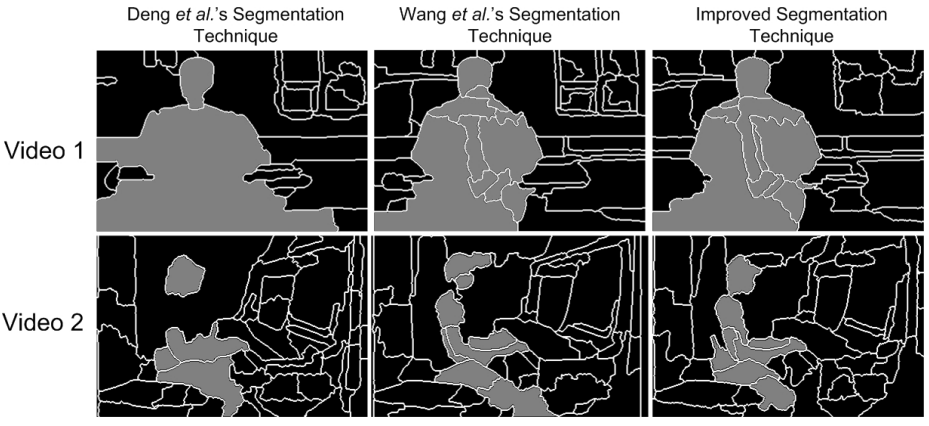

Fig. 3. Segmentation Comparison

4 Conclusions

In this work several improvements were achieved over the original JSEG algorithm in order to allow for a non-parametric clustering of natural scenes and to take advantage of soft-classification maps. The algorithm was also extended in order to improve the region merging process and to track key regions throughout a sequence for the purpose of creating a motion capture system. Results have shown the incredible adaptability of the technique without the need to impose constraints on either the target or its environment thus allowing the technique to be used more efficiently in practical applications.

References

1. Sun, S., Haynor, D.R., Kim, Y.: Semiautomatic Video Object Segmentation Using VSnakes. IEEE Trans. on Circuits and Systems for Video Technology 13(1), 75–82 (2003)
2. Swain, M.J., Ballard, D.H.: Indexing Via Color Histograms. In: Proc. 3rd Intl Conf. on Computer Vision, pp. 390–393 (1990)
3. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. Intl Journal of Computer Vision 1(4), 321–331 (1987)
4. Stauffer, C., Grimson, W.E.L.: Adaptive Background Mixture Models for Real-Time Tracking. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 246–252 (1999)
5. Atev, S., Masoud, O., Papanikolopoulos, N.: Practical Mixtures of Gaussians with Brightness Monitoring. In: Proc. 7th IEEE Intl Conf. on Intelligent Transportation Systems, pp. 423–428 (2004)
6. Hernandez, S.E., Barner, K.E.: Joint Region Merging Criteria for Watershed-Based Image Segmentation. In: Proc. Intl Conf. on Image Processing, vol. 2, pp. 108–111 (2000)
7. Tsai, Y.P., Lai, C.-C., Hung, Y.-P., Shih, Z.-C.: A Bayesian Approach to Video Object Segmentation via Merging 3-D Watershed Volumes. IEEE Trans. on Circuits and Systems for Video Technology 15(1), 175–180 (2005)

8. Wang, D.: Unsupervised Video Segmentation Based on Watersheds and Temporal Tracking. IEEE Trans. on Circuits and Systems for Video Technology 8(5), 539–546 (1998)
9. Chen, J., Pappas, T.N., Mojsilovic, A., Rogowitz, B.E.: Adaptive Perceptual Color-Texture Image Segmentation. IEEE Trans. on Image Processing 14(10), 1524–1536 (2005)
10. DeMenthon, D.: Spatio-Temporal Segmentation of Video by Hierarchical Mean Shift Analysis. University of Maryland, Tech. Rep. (2002)
11. Deng, Y., Manjunath, B.S.: Unsupervised Segmentation of Color-Texture Regions in Images and Video. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(8), 800–810 (2001)
12. Wang, Y., Yang, J., Ningsong, P.: Synergism in Color Image Segmentation. In: Zhang, C., W. Guesgen, H., Yeap, W.-K. (eds.) PRICAI 2004. LNCS (LNAI), vol. 3157, pp. 751–759. Springer, Heidelberg (2004)
13. Georgescu, B., Shimshoni, I., Meer, P.: Mean Shift Based Clustering in High Dimensions: A Texture Classification Example. In: Proc. IEEE Intl Conf. on Computer Vision, pp. 456–463 (2003)
14. Comaniciu, D., Meer, P.: Robust Analysis of Feature Spaces: Color Image Segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 750–755 (1997)
15. Haris, K., Efstratiadis, S.N., Maglaveras, N., Katsaggelos, A.K.: Hybrid Image Segmentation using Watershed and Fast Region Merging. IEEE Trans. on Image Processing 7(12), 1684–1699 (1998)
16. Withers, J.A., Robbins, K.A.: Tracking Cell Splits and Merges. In: Proc. IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 117–122 (1996)