Real-Time Acrobatic Gesture Analysis

Ryan Cassel[1], Christophe Collet[2], and Rachid Gherbi[1]

[1] LIMSI-CNRS, Université de Paris XI, BP 133, 91403 Orsay cedex, France
[2] IRIT, Université Paul Sabatier, 31062 Toulouse cedex, France
{cassel, gherbi}@limsi.fr, collet@irit.fr

Abstract. Gesture and motion analysis is a highly needed process in the athletics field. This is especially true for sports dealing with acrobatics, because acrobatics mix complex spatial rotations over multiple axes and may be combined with various postures. This paper presents a new vision-based system focused on the analysis of acrobatic gestures of several sports. Instead of classical systems requiring modelizing human bodies, our system is based on the modelling and characterization of acrobatic movements. To show the robustness of the system, it was successively tested first on movements from trampoline, and also in other sports (gymnastics, diving, etc.). Within the system, the gestures analysis is mainly carried out by using global measurements, extracted from recorded movies or live video.

1 Introduction

In computer vision systems, techniques of motion analysis are increasingly robust and beginning to have an impact beyond laboratories. Such systems could be useful in order to evaluate the performance of gestures in many applications. We separate communication gestures and sports gestures. These two fields of application use similar algorithms. In the context of communication gesture analysis many studies are devoted to sign language recognition. This task is well known to be hard to perform both in the hand/body/face tracking and recognition processes and in the analysis and recognition of the signs [3][9]. Other gestural studies deal more extensively with topics such as recognition of human activities [1]. The athletics sector is in strong demand for movement analysis. The advent of video techniques in this field already assists users because the video doesn't disturb sport gestures (it is non intrusive techniques). However, there are only few tools allowing automatic analysis in real time, while this type of analysis is necessary for many live sport performances. This paper addresses some of the representative publications on automatic systems of sporting gesture analysis. Yamamoto [4] presents a qualitative study about sporting movement (skiing). The aim of his study is to discriminate movements performed by people classified from novice to experts. Another work by Gopal Pingali [8] shows a system dealing with real time tracking and analysis of a tennis ball trajectory. This system was used for the tennis US Open 2000. The interest of such systems is obvious for many sports. But building them is a real challenge. The sporting context is very constraining from the environmental and gestural points of

S. Gibet, N. Courty, and J.-F. Kamp (Eds.): GW 2005, LNAI 3881, pp. 88–99, 2006.

view. Our study addresses acrobatic gestures. These gestures present a great diversity of complex movements. Each sport based on acrobatics requires a refined analysis in order to improve or judge the gesture quality, and this type of analysis is not commonly available. Sport organizations look forward to the development of such systems, and they agree with the idea that these systems could be a useful complement for training and a helpful tool for judges during competitions. The use of video techniques (non intrusive techniques) derives from specifications set by practice conditions. Thus, the athlete will be never constrained by the system, by contrast with systems involving active sensors for example. Many studies dealing with gesture analysis use sensors placed on the body to extract placement coordinates efficiently. Such systems are known as intrusive, they disturb the human movement making them less natural and they require a complex installation. Acrobatics make the use of sensors problematic. Capture devices such as sensors are expensive, invasive and constraining. However they are very precise compared to image processing. This paper addresses the development and assessment of an analysis system focused on acrobatic gestures. The system uses a fixed monocular passive sensor. The adopted approach is based on movement characterization initially used in trampoline competition. This movement model was improved and is briefly presented in section 2. The architecture system and its corresponding algorithms are described in section 3. Then, section 4 show results issued from evaluation of the system's robustness. Section 5 presents the use of the system in real situations of gymnastic gestures. Finally, a conclusion and some prospective comments are developed in the two last parts.

2 Model of Movement

The system design is based on a model of movement established and used by trampolinists (for more details refer to the *Code of Points* of the Fédération Internationale de Gymnastique [11]). It is based on chronological and axial movement decomposition. The model divides the movement into three parts. The first part relates to the quantity of transversal rotations of the body. The second relates to longitudinal rotations of the body distributed by the quantity of transversal rotations (see example below). The third relates to the body posture. This model of movement leads to a numerical notation described in more detail in [2]. This notation is built as follows:

$$s\ q\ v_1 \dots v_n\ p$$

$$s = [b|f]$$

$$q = [0-9]^*$$

$$v_i = [0-9]^*$$

$$p = [o| < |/]$$

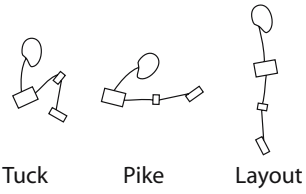

Fig. 1. Human body's positions in trampoline

where s indicates the direction of the figure (f, forward or b, backward), q describes the number of somersaults, by quarters, $v_1 \dots v_n$ represent the distribution and quantity of half twists in each somersault and finally p describes the shape of the figure (tuck = o, pike =< or layout = / figure 1). Thus, b 8 0 0 o describes a double backward tuck somersault (720 of rotation) with no twist whereas f 4 1 / describes a forward layout somersault (360 of rotation) with a half twist.

This notation is valid for all acrobatic activities because it makes possible to identify any acrobatic movement. The recognition of each part of this notation informs us about the quality of realization. It is the basis of our analysis system. Section 5 shows some examples of use of such a system in gymnastics practice (trampoline).

3 System Architecture

The system comprises several connected modules in a hierarchical way. A lower layer level extracts relevant information (typically, pixels of the acrobat). A higher level layer transforms this information into interpretable data and analyzes this data. We present these various layers here.

3.1 Lower Level

The lower level layer extracts pixels of the acrobat. Our approach is to build a statistical model of background image and then to use image subtraction to emphasize the moving elements. We build an original method to eliminate noise, we called it *Block filtering*. And to optimize the acrobat extraction, we use Kalman filtering.

Background Model. In this paper, the term background refers to the pixels which are not moving. Thus, as far as the system is concerned, there is only one person in field of the camera. The gymnast is always moving while the background remains constant. However, in training or competition conditions the background is never fixed. Light variations and people passing behind the athlete make a background image vary. To adapt the background image variations, we use an adaptative generation of background. This generation is based on the luminance

mean of the N last images. The mean is calculated as $m_{(x,y)} = \frac{S_{(x,y)}}{N}$ where $S_{(x,y)}$ is the sum of pixel values in the location (x, y) and N is the number of the last frames collected. Subtraction between the current image and the background highlights fast parts moving. Many pixels are not belonging to the acrobat and are classified as noisy pixels. A filtering must be applied.

Block Filtering. drops noisy pixels. We include in noisy pixels, any moving pixel witch is not belonging to the gymnast. The acrobat is in the foreground of the camera recorder. The subtraction presented above leads to a binary image which contains pixels from the gymnast and pixels belonging to other moving objects. The perspective makes the person in the foreground bigger than other moving objects. Block filtering keeps only the biggest components of the binary image. The size of elements to be dropped is gauged by the size of blocks. To complete this filtering, the image is divided into blocks of size $(n \times m)$ (a grid of size $n \times m$ is applied on the binary image, it will be called *grid block*). Each block is marked as valid, invalid or adjacent. After computing all blocks, the system keeps only valid blocks and adjacent blocks. Invalid blocks are dropped. A block is marked as valid when the proportion of binary pixels is superior to a certain threshold. In the other case it is marked as invalid except if an adjacent block is valid (in a 8-connexity). In this case, the block is marked as adjacent. By defining the block size as larger than the noisy pixels elements' size (moving persons in the background for example), and by defining a threshold greater than the noise elements, the noisy pixels are dropped. Figure 2 shows an example of the algorithm. The upper left image corresponds to the original image. The upper right image corresponds to the result of the subtraction operation. It shows many pixels belonging to the noise created by elements moving in the background. The result of block filtering is given in the last image.

Let C be the mathematical expression for a block at pixel (x, y) on the grid block :

$$C(x,y) = \frac{\sum_{i=n.(x-1)}^{n.((x-1)+1)} \sum_{j=m.(y-1)}^{m.((y-1)+1)} I_{bin}(i,j)}{n.m}$$

where $I_{bin}(i, j)$ is the binary image at pixel (i, j). The criteria for selecting valid, invalid and adjacent block are defined as folow :

$$Block(x,y) = \begin{cases} \text{Valid}, & \text{if } C(x,y) \geq Th; \\ \text{Adjacent}, & \text{if } Block(x+i, y+j) = Valid \; \forall i \in [-1;1], \forall j \in [-1;1]; \\ \text{Invalid}, & \text{otherwise.} \end{cases}$$

where $Block(x, y)$ is a block at pixel (x, y) and Th the percentage of pixels wich should fill the block (fixed by user).

This filter is efficient when there is enough difference between foreground and background. However, a collective movement of the audience like a "holla" strongly disturb the algorithm. In this paper, the corpus used for this study does not include such situations.

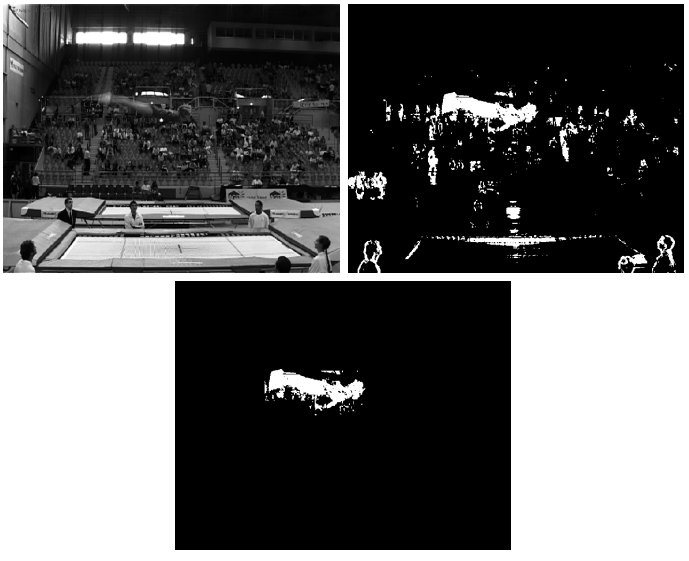

Fig. 2. Original image (top left), Binary image (top right) and Block Filtered image (bottom)

Kalman Filter. To improve time processing, it is important not to compute the entire images. Only the acrobat is relevant. Consequently, it is necessary to track the acrobat. Kalman filtering allows the system to predict and estimate displacement [5].

The region of interest of the image is reduced to a box around the gymnast (bounding box). All computing tasks are reduced to this bounding box. The Kalman filtering gives efficient prediction of the box's location. In addition to reducing the time processing, the bounding box focuses on the athlete. New elements around the bounding box do not come to disturb processing (for example, a person passing in the background). After an initialisation stage, the system is focused on the gymnast.

3.2 Motion Analysis (High Level)

Our characterization defines acrobatics with axial descriptions and with human body shape. Motion analysis part first extracts body axis and then analyses the body shape. Mathematical calculation gives the axis, and surface analysis gives the shape. These data leads to a part of the numerical notation [2].

Determination of Rotational Quart - Calculation of 2D Orientation. From the binary image and according to the bounding box, the system computes the principal axis of the binary shape. As described in [6] we use a mathematical method to determine the athlete axis. For discret $2D$ image probability distributions, the mean location (the centroid) within the search window, that is computed at step 3 above, is found as follows:

Find the zeroth moment :

$$A = \sum_x \sum_y I(x, y).$$

Find the first moment for x and y :

$$M_x = \sum_x \sum_y xI(x, y), \quad M_y = \sum_x \sum_y yI(x, y).$$

Mean search window location (the centroid) then is found as

$$\bar{x} = \frac{M_x}{A}, \quad \bar{y} = \frac{M_y}{A}.$$

The 2D orientation of the probability distribution is also easy to obtain by using the second moments in the binary image, where the point (x,y) ranges over the search window, and I(x,y) is the pixel (probability) value at the point (x,y).

Second moments are :

$$M_{xx} = \sum_x \sum_y x^2 I(x, y), \quad M_{yy} = \sum_x \sum_y y^2 I(x, y), \quad M_{xy} = \sum_x \sum_y xyI(x, y).$$

Let

$$a = \frac{M_{xx}}{A} - \bar{x}^2, \quad b = 2\left(\frac{M_{xy}}{A} - \bar{x}\bar{y}\right), \quad c = \frac{M_{yy}}{A} - \bar{y}^2.$$

Then the object orientation, or direction of the major axis, is

$$\theta = \frac{\arctan\left(\frac{b}{a-c}\right)}{2}.$$

The first two eigenvalues, that is, length and width, of the probability distribution of the blob found by the block filtering may be calculated in closed form as follows:

Then length l and width w from the distribution centroid are

$$l = \sqrt{\frac{(a+c) + \sqrt{b^2 + (a-c)^2}}{2}}, \quad w = \sqrt{\frac{(a+c) - \sqrt{b^2 + (a-c)^2}}{2}}.$$

When used in human tracking, the above equations give body roll, length, and width as marked in the source video image in Figure 3.

Rotation Tracking. The computed axis leads to the body orientation but it is not oriented. The given orientation is available at $\pm\pi$. Indeed, the function arctan is defined on $]-\frac{\pi}{2};\frac{\pi}{2}[$. We use biomechanical constraints to eliminate ambiguous measurements. A study on physical constraints in acrobatics defined maximum angular velocities for twist and somersaults. The maximum velocity for somersault is $\omega_{max} = 22\ rad.s^{-1}$ (or $\omega_{max} = 0.89\ rad/image$ for videos running at 25 $images/s$). Variations around π are physically impossible. Thus the instantaneous angular velocity is : $\omega = \theta_{t-1} - \theta_t\ [\pi]$. And the correct orientation θ' is : $\theta'_t = \theta'_{t-1} + \omega$. This is a relative orientation depending on the first orientation.

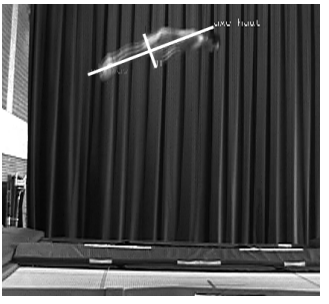

Fig. 3. Human body's orientation, length and width

Quarters Detection. Find correctly the number of quarters of somersaults is not so obvious. When a gymnast execute a somersault, it is frequent that the body axis does not describe a entire rotation for a simple somersault (for a f 4 0 o the total rotation is $\theta'_t < 2\pi$). However, the system has to detect 2π i.e. 4 quarters of transversal rotation because the acrobat starts from feet and arrives on feet (figure 4). For example, a backward somersault (b 4 0 o in clockwise) starts at $\frac{\pi}{4}$ and finishes at $\frac{7\pi}{4}$. The number of quarters is 3 whereas the system has to detect 4. We introduce 3 methods to detect quarters. The first method is to calculate average angular velocity on the portions of somersault where this velocity is constant. A null angular acceleration generates a constant speed. While calculating the mean velocity of the constant phase one manages to have a good idea of the salto carried out. By deferring the mean velocity on the unit of the jump, we compensate the orientation problems of departure and the end of the jump. Another method is to compare the takeoff orientation with the landing orientation. This gives an exact measure. The last method consists in counting the number of dial crossed by the axis of the body. Somersault rotations are represented by a disc cut out in dials. This disc is divided into four zones ($[0; \frac{\pi}{2}[$, $[\frac{\pi}{2}; \pi[$, $[\pi; \frac{3\pi}{2}[$, $[\frac{3\pi}{2}; 2\pi[$). To each time the axis enters a new dial it is entered. These three methods lead to the right detection. The differences among three methods of detecting quarters are : The first method gives sometimes too quarters due to the inertia of fast somersaults. However the second method is

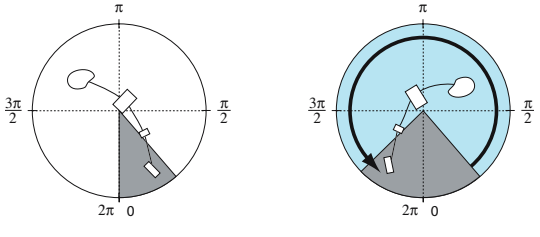

Fig. 4. Quarters rotation determination

concerned with the problem of the number of effective quarter and thus gives less quarters. The third method is correct around ± 1 quarter (when the acrobat starts from his back, he is hidden by the trampoline, this lead to miss quarter). By calculating the average of the three methods, quarters are correctly detected (table 1).

Table 1. Evaluation of recognition and tracking algorithm

Type	**Test**	**Evaluation**
Tracking correlation	93 %	94 %
Tracking standard deviation	26 pixels	26 pixels
Rotations	100 %	98 %
Positions o	57 %	50 %
Positions <	50 %	54 %
Positions /	100 %	100 %

Position Evaluation. The evaluation of the body shape, leads to *tuck*, *pike*, or *layout*. To this goal, the system needs the axis l, w and the surface of the gymnast (wich are variable size during somersault). The average l_m of the axis l and w is the diameter of a disc C centred on the centroid of the acrobat. The more the acrobat is tuck, the more the ratio of w on l is close to 1. And the more the ration of the surface of the disc C and the surface gymnast is close to 1. Most of the acrobat's pixels is included in the disc C. When these results are close to 0, we can conclude that the body shape is layout. This first stage makes it possible to differentiate tuck from layout. Nevertheless, tuck and pike are similar shapes. The pike shape is less compact than the tuck shape. Currently, the system is not able to make the difference between these two body shapes. Because acrobats are not perfect, these body shapes are not carried out perfectly. Compared to the rotational quarters' part, we have a similar problem and we have to discriminate right shapes.

Twist Evaluation. Twists are not detected yet. The recognition of all elements in the numerical notation is not complete. The twist detection is undoubtedly the most complex part to realize because this movement mix both transversal rotations and longitudinal rotations.

4 Experimental Result

To assess our system we carried out a video corpus which we manually labelled in part. We pinpointed the position of the head, hands, base, knees, and feet. The video corpus comprises more than 100 sequences and more than 1000 figures performed by 7 athletes. We divided the corpus into two parts: one for the algorithms adjustment and the other to evaluate them. Evaluations presented below were carried out on sequences with no audience in the background.

Body Tracking. From our labelled data, we deduced the centroid of the athlete (centroid of all labelled body parts). We calculated the correlation between the data measured manually and the data calculated automatically by the system. The result is presented in the table 1.

Rotationnal Quart Recognition. Quarters were counted manually and the counts were then compared with those calculated automatically. The percentage in table 1 represents the relationship between the quarters manually counted and those recognized by the system.

Position Recognition. In the same way, each body shape was evaluated. They were first manually labelled and the results were then compared with those recognized by the system.

Tracking results are sufficient to obtain a good recognition of the somersaults. The detection of quarters depends on the environment. The system limits appear clearly. Indeed it is not yet possible to use this system in a competition context, with a frantic audience in the background! In that situation, the recognition rate decreases shortly. A better primitives extraction algorithm is then required. The results for the recognition of quarters show that the system is effective, despite being occasionally mistaken in extreme cases. Recognition of the body shape has not been successful so far because it is not possible yet to discriminate between tuck and pike. The system regularly confuses the two shapes because they are relatively similar. However information which the system extracts already makes it possible to be exploited. Under training conditions the system shows a very good robustness. We will see in the section 5 that it was used in real conditions to adapt the sporting training.

As for processing time, the system runs in real time. After an initialization stage, one image is completed on average in 0.019 seconds on a 2.6 Ghz PC. The real time allows coaches to effectively use this system.

5 Training Evaluation

The system suggested in this article is not finalized. However it is able to extract significant information from training. In gymnastic apparatus such as the bars (high bar and uneven bars), the rings, the trampoline, the beam, the system brings many information which are not obvious during training. The system helps the coach on not easily remarkable information. We present an example during trampoline training.

We use the system to adapt the trampoline training. During competitions preparation, coaches prepare with gymnasts, sequences of 10 elements which will have to be perfect. An element is a jump, a skill (a salto for example). The ideal one is to carry out these 10 elements with a constant height. Figures 5 and 6 illustrate the amplitude variations of the 10 elements of a sequence. Each element is described by using the numerical notation quoted previously. The development of these sequences is not simple. Each gymnast has its characteristics and its facilities in realising elements. A movement can be appropriate to an athlete but

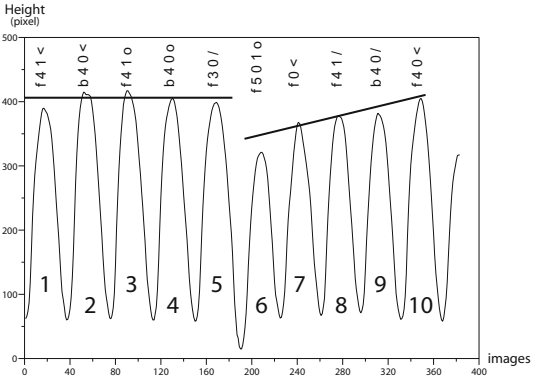

Fig. 5. Initial sequence (not adapted)

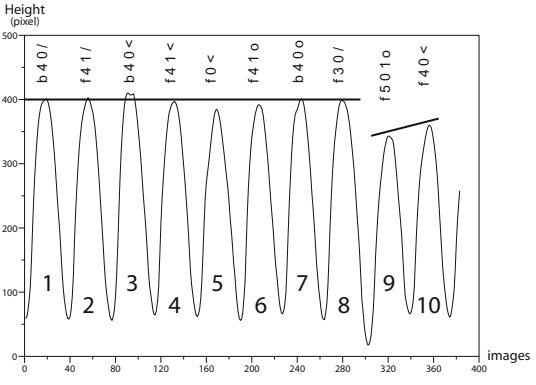

Fig. 6. Corrected sequence (adapted)

not for another. The example below illustrates a traditional sequence carried out by many gymnasts. During training, this sequence passes with difficulty for an individual. The gymnast has difficulties practising the movement. The traditional error would be he has to practice many more. However by looking at the figure 5 one notes a first relatively constant part (elements from 1 to 5), a setback (with the 6th element) and an increase towards the initial height until the end of the movement. The connection 5th element, 6th element is too difficult to realize at this place of the sequence. The idea is to move this connection at the end of the sequence. The system evaluates the sequence again.

In the corrected movement (figure 6) one finds this abrupt loss of amplitude (element 8 to 9). But this difficulty does not interfere any more with the other elements.

The athlete is an actor and can not see what is going wrong; it is the role of the coach. This one cannot be attentive with all details, especially when they are not easily locatable. The system highlights a considerable loss of amplitude which leads to a sequence mediocrity. The accused element then is replaced or moved.

The system evaluates the amplitudes variations on this example. It is also able to evaluate the velocities and acceleration of translation and angular rotations of the global human body. This information is not easily reachable by the human eye. In the next improvements, the system should be able to recognize and evaluate them.

6 Perspectives and Conclusion

The first prospect for the system is to be able to evaluate twists quantitatively. The use of optical flow could solve this problem. The first measurements of the optical flow [7] leads us to pursue our investigations. The optical flow remains an extremely expenditure in computing time and could offset the value of real time computing. We therefore propose to calculate the optical flow on parts of the bounding box after having restored the axis of the body to a vertical position. The rotational component would be cancelled out following the transversal axis and the translational component. This calculation should highlight only the longitudinal rotational component. The second prospective element is to finalize the system to make it a robust recognition system for acrobatic gesture.

This paper presents an analysis system of sporting gestures by global measurements. The system does not identify the parts of the human body but bases its recognition on measures of the global human body. Such a system is taken out of laboratories because it makes analysis in sport context and not in laboratories context. The system is not intrusive. There are no constraints for the sportsman (no sensor obstructing), it preserves the naturality of the analyzed gesture. The simplicity of implementation and the low cost of the hardware make the system accessible to sports coaches. Global measurement can lead to a robust recognition thanks to an adequate characterization. The low complexity of the algorithms allows real time. The system gives useable results and it is already used for training. The system reaches its limits when the camera is not fixed or when a crowd of people is moving in the background. In the same way the system does not allow analysis of two persons in the field of the camera. The system recognizes acrobatic gesture by global measurement. The terminology employed is effective and is recognized in the international trampoline community. It is certainly not very pleasant but each element has a translation in every language. This kind of system is very helpful for coaches and judges in competition.

References

1. H. Lakany and G. Hayes, "An algorithm for recognising walkers," *Second IAPR Workshop on Graphics Recognition*, Nancy, France , pp. 112-118 , August 22-23, 1997.
2. Ryan Cassel and Christophe Collet, "Tracking of Real Time Acrobatic Movements by Image Processing," *5th International Gesture Workshop*, Genova, Italy, pp. 164 - 171 , April 15-17, 2003.

3. A. Braffort, A. Choisier, C. Collet, et al., "Toward an annotation software for video of Sign Language, including image processing tools and signing space modelling," *4th International Conference on Language Resources and Evaluation*, Lisbonne, Portugal, 2004.
4. Masanobu Yamamoto, Takuya Kondo, Takashi Yamagiwa and Kouji Yamanaka, "Skill Recognition," *3rd IEEE International Conference on Automatic Face and Gesture Recognition,* pp.604-609, 1998.
5. Greg Welch, Gary Bishop, "An Introduction to the Kalman Filter," *TR 95-041*, Technical Report, Department of Computer Science, University of North Carolina, NC, USA, 2002
6. Ryan Cassel and Christophe Collet, "Caracterization and Tracking of Acrobatic Movements for Recognition," *Workshop in 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, France, 2004.
7. Changming Sun, "Fast Optical Flow Using 3D Shortest Path Techniques," *Image and Vision Computing*, Vol. 20, no.13/14, pp.981-991, December, 2002.
8. Gopal Pingali, Agata Opalach, Yves Jean, "Ball Tracking and Virtual Replays for Innovative Tennis Broadcasts," *International Conference on Pattern Recognition (ICPR'00)*, p. 4152, Volume 4, September 03 - 08, Barcelona, Spain, 2000.
9. Fanch Lejeune, Annelies Braffort and Jean-Pierre Descls, "Study on Semantic Representations of French Sign Language Sentences," *4th International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, P. 197-201, London, UK, 2001.
10. Kong Man Cheung and Simon Baker and Takeo Kanade, "Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June, 2003.
11. Fédération Internationale de Gymnastique, "Trampoline Code of Points 2005," www.fig-gymnastics.com.