Video Annotation by Motion Interpretation using Optical Flow Streams[†]

G. Sudhir[*] and John C.M. Lee[*]

Technical Report HKUST-CS96-16
July 13, 1997

[*]Department of Computer Science
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

[†]This research was supported by Sino Software Research Center of Hong Kong University of Science and Technology by grant SSRC94/95.EG08.

GS is a research associate.

Abstract

A new approach to automatic annotation of video sequences by dominant motion interpretation is presented. Unlike others, we separate the optical flow into two categories - *singular* and *non-singular* - which as we show is a more natural way of classification for the purpose of *dominant* motion interpretation. We show that identification of patterns created by such natural categories, which can be observed from the *measured* optical flow, can help focus the interpretation of *dominant* motion in video segments. For robust observation of such natural patterns, we propose the computation of optical flow streams (OFS) from the video data and analyse the OFS for extraction of dominant motion content in the video segments. Our proposed approach has both bottom-up and top-down schemes suitably applied. The bottom-up scheme computes the OFS purely by local optimization of optical-flow equation. Then, the top-down scheme interprets the signature of the projection of the OFS on to the image coordinates for the detection of the natural category of the observed flow. Finally, further bottom-up analyses are done for sub-classification of the motion content in the video segments. The advantage of the proposed approach is robustness in the extraction of dominant motion content in a video segment. We demonstrate this on a variety of real video sequences by generating the automatic motion annotation of the video frames and comparing with manual motion annotation.

1 Introduction

With the advent of video technology, video sources and video data have become a common place. Almost everything from entertainment to education is recorded as video data and gets distributed across the globe. This has lead to an enormous increase in video data which has become extremely difficult to manage for the sheer volume of it, thus making visual information management a grave necessity and a major topic in information technology.

Making efficient use of video data requires that the data be stored in an organized way. For this, as much as possible, it must be associated with appropriate *semantic* indices in order to allow any future retrieval. This has been called *content-based retrieval* in literature [1, 2]. Traditional database management techniques are only capable of managing data in alphanumeric form, and hence, in the past, image and video data have been managed using *keywords*, which are either metadata or brief descriptions of images [3]. However, this is extremely laborious and time consuming, especially for the management of video data, thus calling for research and development of techniques for automatic extraction of *semantic* indices for annotation to enable *content-based* retrieval of video segments from large corpus of video data.

Amongst various such *semantic* indices, color, texture, shape and motion form some of the most necessary ones, for annotation of the video frames to capture the *theme* present in the video segments. Though there have been many research efforts reported in literature in using color, shape and texture [4, 5, 6], mostly in the case of still imagery, there have been not many research results addressing automatic extraction of motion indices especially in the context of structuring video data. This paper focuses on automatic extraction of *dominant* motion content in a video segment.

Motion content in a video segment is a very powerful cue for organising video data because, during retrieval, one can focus only on the most interesting *motion theme* and skip all the uninteresting ones. So, if accurate motion annotation of video frames is done, the searching process can utilize the annotated motion theme in a video segment, along with other *semantic* annotations, for either skipping or for further exploration of the video segment.

For the pupose of motion annotation, automatic identification of *motion theme* becomes necessary. However, motion understanding is a very rich and complicated problem in general and has still been a hot topic of research for the computer vision community [7]. Hence, currently, automatic identification of whether a video segment contains (*i*) object motion only, (*ii*) no motion, or *iii*) camera motion and possible further sub-classification of the nature of camera motion, form some of the important themes for the video segments to be annotated with. Since camera motion can be *pan*, *zoom*, *tilt*, *translation* or any combination of these in general, identification of the *dominant* camera motion nature is important for generation of useful motion annotations.

Though not in the context of automatic video annotation, several authors have examined the problem of extraction of camera motion parameters. Most of the literature is in the context of motion estimation and compensation for video compression and coding [8, 9, 10, 11]. Most of the literature available in the context of video compression and coding try to ignore the camera translations, with an assumption of small camera motion between adjacent frames. Srinivasan *et al* [12] report the estimation of horizontal and vertical camera translations utilizing the fact that those components are parallel in the image. However, they work under the assumption that there is purely camera motion and any object motion present will affect their procedure. Also, for simplicity, they ignore the camera *Z-translation* terms in the equations of the optical flow (see Section 2) altogether. Rangarajan and Shah [13] present an analytical approach for the study of dynamic events analyzing the locus of focus of expansion (FOE) of a group of motion trajectories. They show that qualitative inference of dynamic motion events can be determined from such locus of FOE. However, they assume that FOE has been already computed accurately and they do not report results regarding the robustness of implementation of their scheme in the context of automatic video annotation. Akutsu *et al* [14] address video indexing using motion vectors wherein they report a Hough space analysis for the extraction of *dominant* camera motion parameters. However, they ignore the camera translations in their analysis.

Importantly, all these research works, do not utilize the power of patterns, created by the type of camera motion, in the observed optical flow, for the possible simplification of many problems in extracting *dominant* camera motion information. Furmüller [15], reports a method to look for such patterns, which depend only on subsets of motion parameters, in the context of estimation of 3D camera motion parameters and even works from the assumption that only *normal flow* (see [15]) has been computed. Eventhough, the suggested approach is good for the cases when the estimation of all (or most) of the camera motion parameters is required, he reports a relatively complicated pattern search and fitting technique, which depends upon the accurate measurement of image gradients from the observed image data, for the separation of rotational and translational components. This may not be necessary for the purpose of extraction of *dominant* motion content in a video segment, which forms the main kind of information with which the video frames need to be annotated to support *content-based* retrieval using the motion *theme*.

With *dominant* motion annotation as the main objective, in this paper, we take a new and different approach to look for patterns which as we show are more naturally observed in the measured optical flow. In fact, unlike Fermüller, we do not want to separate the translational and rotational components, rather we separate them to *singular* and *non-singular* categories (see Section 2 for more details). Moreover, we introduce what are called optical flow streams (OFS) (see Section 3.2) for robust detection of such *singular* and *non-singular* patterns. We exploit the observation of such natural patterns for our purpose of simplification as well as robust extraction of *dominant* motion in a video segment for annotation. However, our approach has similar effect of simplifying the (3D) camera motion parameter estimation problems also. This has relevance to accurate estimation camera motion parameters and motion compensation for video data compression. In this paper, we concern ourself to robust extraction of *dominant* motion content in video segments for motion annotation.

The rest of the paper is organised as follows. In Section 2, we classify the motion field observed on image into two categories and describe our approach to interpret the *dominant* motion content in video segments. In Section 3, we describe our method of measurement of motion field (rather, its approximation) by the computation of optical flow field and OFS. In Section 4, we give our complete algorithm for automatic motion annotation of the video segments. In Section 5, we present experimental results on a variety of video data to demonstrate the robustness of our proposed motion annotation scheme. Finally, in Section 6, we summarize the contributions of the paper as well as some future directions for knowledge-based motion annotation schemes using the OFS signatures and conclude the paper.

2 Motion Classification and Interpretation

In this section, we present our approach for classification of *dominant* camera motion into different categories based on the observation of optical flow field on the image. For the analysis in this section, we assume that it has already been detected that camera motion has occured. In the Section 3, we describe a method for detection of such camera work in a video segment as against only object motion or no motion.

For clarity of presentation, we follow the standard way of analysis using the camera centered coordinate system. Figure 1 shows the camera geometry and the perspective projection mechanism. Let f_x and f_y be the focal lengths of the camera in x and y directions, respectively. Note that the 3D world coordinates are written in capital letters ($\mathbf{P} = (\mathbf{X}\,\mathbf{Y}\,\mathbf{Z})^{\mathbf{t}}$) where as the image coordinates are written in small letters ($\mathbf{p} = (\mathbf{x}\,\mathbf{y})^{\mathbf{t}}$).

The perspective projection of a world point $\mathbf{P} = (\mathbf{X}\,\mathbf{Y}\,\mathbf{Z})^{\mathbf{t}}$ onto the image plane is given by $\mathbf{p} = (\mathbf{x}\,\mathbf{y})^{\mathbf{t}}$ where

$$x = f_x \frac{X}{Z}, \quad and, \quad y = f_y \frac{Y}{Z}. \tag{1}$$

Let the camera motion be given by

$$\mathbf{V} = \mathbf{T} + \mathbf{\Omega} \times \mathbf{P} \tag{2}$$

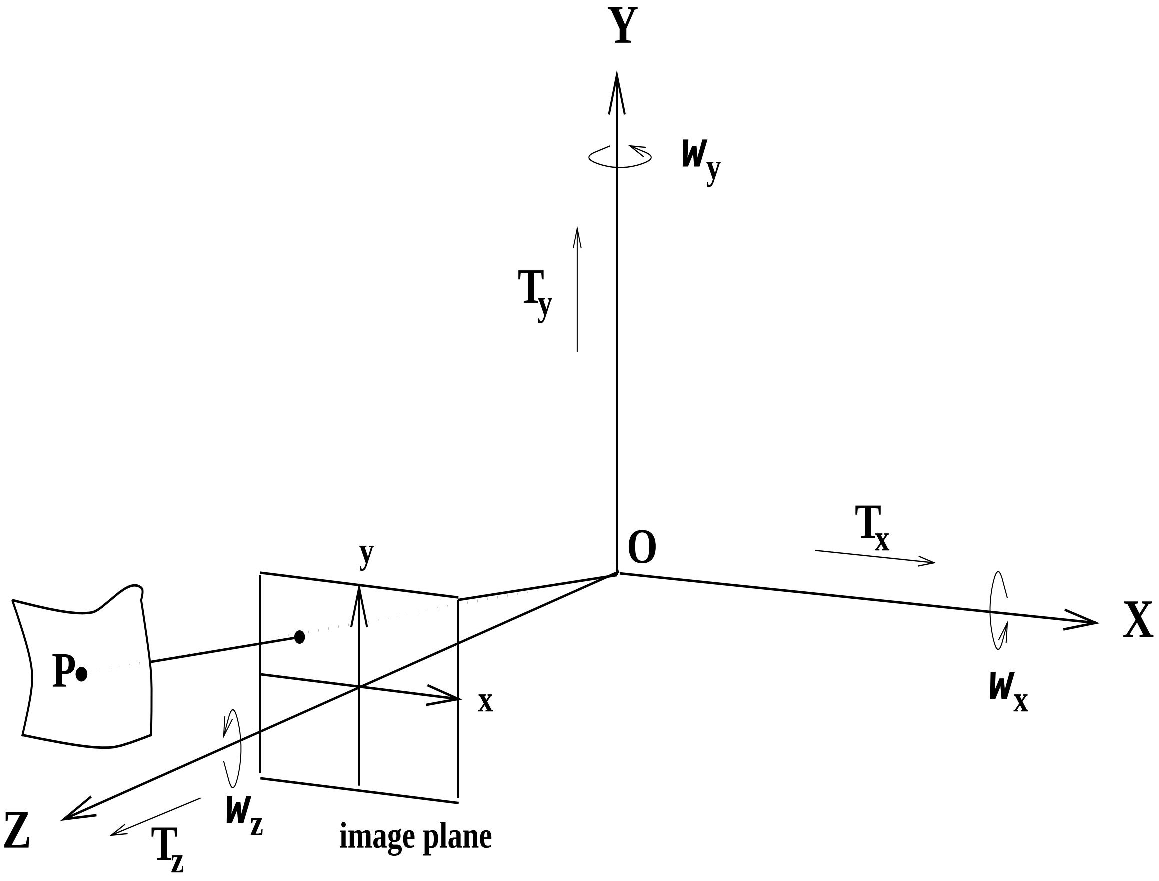

Figure 1: Camera Geometry

where, $\mathbf{V} = (\mathbf{V_X}\, \mathbf{V_Y}\, \mathbf{V_Z})^{\mathbf{t}}$ is the 3D velocity vector of the point whose position vector is $\mathbf{P} = (\mathbf{X}\,\mathbf{Y}\,\mathbf{Z})^{\mathbf{t}}$, $\mathbf{T} = (\mathbf{T_X}\,\mathbf{T_Y}\,\mathbf{T_Z})^{\mathbf{t}}$ is the 3D vector to model the camera translation, and, $\mathbf{\Omega} = (\omega_{\mathbf{X}}\, \omega_{\mathbf{Y}}\, \omega_{\mathbf{Z}})^{\mathbf{t}}$ is the 3D vector to model the angular velocity of camera rotation. Note that $\mathbf{A}^{\mathbf{t}}$ denotes the transpose of $\mathbf{A}$ and that the above camera motion model has a total of 6 parameters, 3 each in $\mathbf{T}$ and $\mathbf{\Omega}$.

Let $\mathbf{OF}(\mathbf{x}, \mathbf{y}) = (\mathbf{u}(\mathbf{x}, \mathbf{y})\, \mathbf{v}(\mathbf{x}, \mathbf{y}))^{\mathbf{t}}$ be the observed optical flow at location $\mathbf{p} = (\mathbf{x},\, \mathbf{y})^{\mathbf{t}}$. From the perspective projection model (Eq. 1) and the camera motion model (Eq. 2), it is easy to show by differentiation that

$$\begin{aligned} u(x,y) &= -\left(f_x + \frac{x^2}{f_x}\right)\omega_Y + \frac{f_x}{f_y}\, y\, \omega_Z - \frac{f_x}{Z}\, T_X + \frac{xy}{f_y}\, \omega_X + \frac{x}{Z}\, T_Z \\ v(x,y) &= \left(f_y + \frac{y^2}{f_y}\right)\omega_X - \frac{f_y}{f_x}\, x\, \omega_Z + \frac{f_y}{Z}\, T_Y - \frac{xy}{f_x}\, \omega_Y + \frac{y}{Z}\, T_Z \end{aligned} \tag{3}$$

The above equations for the x and y components of optical flow do not include the terms corresponding to camera *zoom* operation. Adding the terms corresponding to camera *zoom* (we borrow the corresponding terms from [16, 12]), we get

$$\begin{aligned} u(x,y) &= -\left(f_x + \frac{x^2}{f_x}\right)\omega_Y + \frac{f_x}{f_y}\, y\, \omega_Z - \frac{f_x}{Z}\, T_X + \frac{xy}{f_y}\, \omega_X + \frac{x}{Z}\, T_Z + \\ &\quad f_x\, tan^{-1}(\frac{x}{f_x})\, (1 + \frac{x^2}{{f_x}^2})\, r_{zoom} \\ v(x,y) &= \left(f_y + \frac{y^2}{f_y}\right)\omega_X - \frac{f_y}{f_x}\, x\, \omega_Z + \frac{f_y}{Z}\, T_Y - \frac{xy}{f_x}\, \omega_Y + \frac{y}{Z}\, T_Z + \\ &\quad f_y\, tan^{-1}(\frac{y}{f_y})\, (1 + \frac{y^2}{{f_y}^2})\, r_{zoom} \end{aligned} \tag{4}$$

where, r_{zoom} is the common *zoom* factor in the x and y direction.

The above equations model most of the basic camera motions (operations). It should be noted that there are a total of 9 parameters in the above equations ($\mathbf{T}$, $\mathbf{\Omega}$, f_x, f_y, r_{zoom}), and that the parameters are involved quite nonlinearly. But that is not all; there is an unknown variable Z also involved above. So, even if we have N different measurements ($N \gg 9$) of optical flow $OF(x_i, y_i)$, $i = 1\; to\; N$ on the image, for each such measurement, there will be an unknown variable Z_i introduced extra, and hence the problem of estimation of the above parameters is quite difficult in general, and hence it has occupied considerable research effort by the people in computer vision community [15].

In this context, if we have some *apriori* knowledge of the typical kind of *dominant* camera motion in a video segment, or, better, if we have some way of *infering*, from the measurement data itself, about different classes of dominant camera motions that could have caused the measurement, we will be able to (*i*) possibly reduce the number of parameters required to describe the camera motion in a video segment, and/or (*ii*) linearize the reduced parameter estimation problems, and/or (*iii*) be able to robustly interpret the dominant camera motion in a given video segment and thus be able to reliably annotate the video segment with motion indices.

With the last (third) reason as the motivation for our purpose of video annotation by motion interpretation, we classify the above equations for optical flow (Eq. 4) into two categories as follows:

$$
\begin{aligned}
u(x,y) &= \{u_{singular}(x,y)\} + \{u_{non-singular}(x,y)\} \\
v(x,y) &= \{v_{singular}(x,y)\} + \{v_{non-singular}(x,y)\} \qquad (5)
\end{aligned}
$$

$$(6)$$

where, we define an optical flow term as *singular* if it vanishes at camera center and as *non-singular* if it does not (the motivation for choosing these names comes from Verri *et al* [17]), and,

$$
\begin{aligned}
u_{singular}(x,y) &= \left\{ -\frac{x^2}{f_x}\,\omega_Y + \frac{f_x}{f_y}\,y\,\omega_Z + \frac{xy}{f_y}\,\omega_X + \frac{x}{Z}\,T_Z + \right. \\
&\qquad \left. f_x\,tan^{-1}(\frac{x}{f_x})\,(1+\frac{x^2}{{f_x}^2})\,r_{zoom} \right\} && (7)\\
u_{non-singular}(x,y) &= \left\{ -f_x\,\omega_Y - \frac{f_x}{Z}\,T_X \right\} && (8)\\
v_{singular}(x,y) &= \left\{ \frac{y^2}{f_y}\,\omega_X - \frac{f_y}{f_x}\,x\,\omega_Z - \frac{xy}{f_x}\,\omega_Y + \frac{y}{Z}\,T_Z + \right. \\
&\qquad \left. f_y\,tan^{-1}(\frac{y}{f_y})\,(1+\frac{y^2}{{f_y}^2})\,r_{zoom} \right\} && (9)\\
v_{non-singular}(x,y) &= \left\{ f_y\,\omega_X + \frac{f_y}{Z}\,T_Y \right\} && (10)
\end{aligned}
$$

Remark 2.1 *We treat the terms $\frac{f_x}{Z}\,T_X$ and $\frac{f_y}{Z}\,T_Y$ as conditionally non-singular in the sense that if the value of Z is within a finite range, these terms do not contribute to the qualitative property of singularity as defined above. Though this holds for most indoor video sequences, however, this will not be the case for outdoor video sequences since the value of Z for the sky will be relatively infinite compared to the values of Z for objects in the outdoor scene. We will discuss this again in Section 6.*

It should noted, from the above classification of the optical flow field observed on the image, that more terms contribute to the *singular* flow. Thus, if we can detect, from the measured optical flow, the type of observed dominant flow, we can separate the analysis of each of the two categories of flow: *singular* and *non-singular*. In Section 4, we describe our approach to detect whether a singular optical flow has been observed or not, by a simple and robust technique. To visually depict the *qualitative* property of *singularity* of optical flow, we show in Fig. 2 the typical signatures of OFS magnitudes computed (see Section 3 for more) for different camera work types which are detected correctly by the test. Assuming that such a separation has been done, in the following subsections, we describe our methods for further sub-classification of flow in each category in a systematic manner.

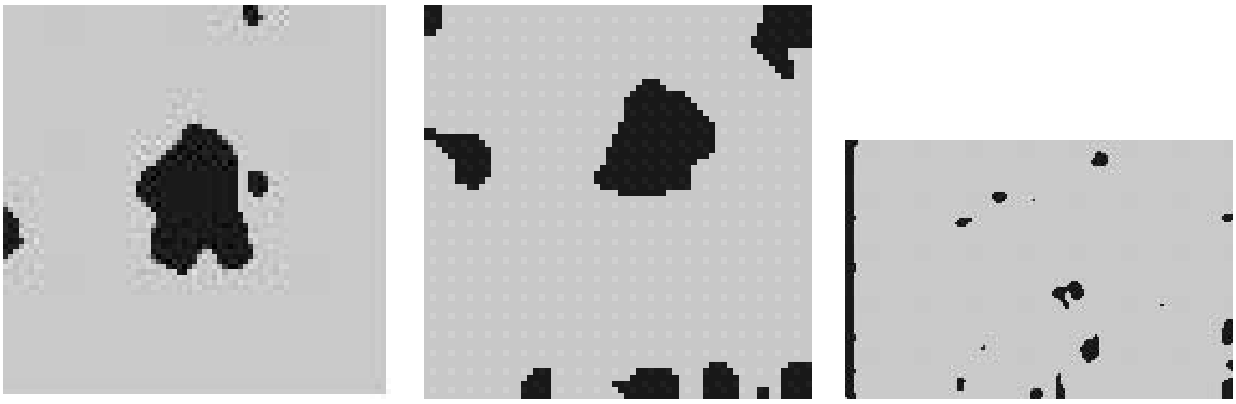

Figure 2: Typical OFS magnitude signatures for different types of camera work: (a) left: that observed on a frame of *Fleet* image sequence (camera *zoom*; see Section 5 for some of the images), (b) middle: that observed on a frame of image sequence shown in Fig. 3 (camera *Z-translation*) and (c) right: that observed on a frame of *Table Tennis* image sequence (camera *pan*; see Section 5 for some of the images). Note that the first two belong to *singular* category and the third belongs to non-singular category.

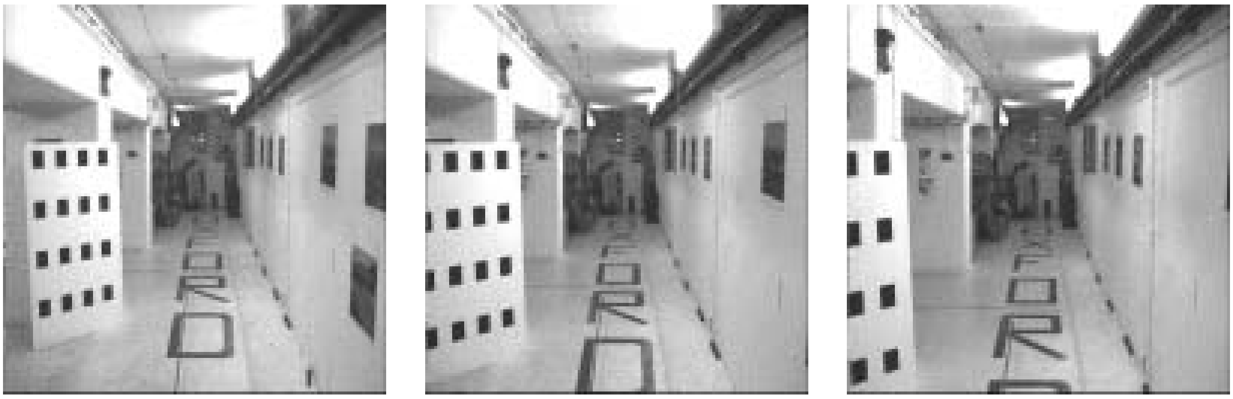

Figure 3: Some frames of a camera *Z-translation* sequence.

2.1 Subclassification of Singular Flow

As can be seen from Eq. 7 and Eq. 9, it is clear that (i) there are no constant (unknown) terms and (ii) there are both linear and non-linear terms involving variables x and y. Also, since the observed flow is classified as *singular*, we can delete the terms $-\frac{x^2}{f_x}\,\omega_Y$ and $\frac{xy}{f_y}\,\omega_X$ in $u_{singular}(x,y)$ and the terms $\frac{y^2}{f_y}\,\omega_X$ and $-\frac{xy}{f_x}\,\omega_Y$ in $v_{singular}(x,y)$ which involve the *tilt* parameter ω_X and the *pan* parameter ω_Y, since, otherwise, these *pan* and *tilt* parameters should have caused the observation of a *non-singular* flow in the first place (see Eq. 7 and Eq. 9; see also Remark 2.1).

A closer observation of the rest of the singular terms indicates much more:

(i) the *Z-rotation* parameter ω_Z is weighted by variable y in $u_{singular}(x,y)$, where as, it is weighted by variable x in $v_{singular}(x,y)$,

(ii) the *Z-translation* parameter T_Z is weighted by variable x in $u_{singular}(x,y)$, where as, it is weighted by variable y in $v_{singular}(x,y)$, and,

(iii) the *zoom* parameter r_{zoom} is weighted by a non-linear function of the variable x in $u_{singular}(x,y)$, where as, it is weighted by similar non-linear function of variable y in $v_{singular}(x,y)$.

From these crucial observations, we design a sub-classification method as follows. We estimate a 2D affine motion model consisting of 6 parameters for the *singular* optical flow as follows:

$$\mathbf{OF(x,y)} = \mathbf{A\,p} + \mathbf{b} \tag{11}$$

where

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \textbf{and} \quad \mathbf{b} = \begin{pmatrix} e \\ f \end{pmatrix}.$$

Here, a, b, c, d, e, f are the 6 parameters. Since the observed optical flow is *singular*, the norm of **b** should be less that a threshold $Th_{singular}$. Hence, most of the information for the *singular* flow is contained in the 2×2 matrix **A**.

Now consider the following two orthogonal bases (of the possible 4 linearly independent bases) for an operator on a 2D vector space.

$$\mathbf{I_1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \textbf{and} \quad \mathbf{I_2} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{12}$$

From the three observations described above and from the above equation, we can easily detect whether the *dominant* camera motion is due to *Z-rotation* or not as follows.

Projection test

> Compute, in the lease square error sense, the 2D affine motion model (**A** and **b**) for the optical flow vectors in the whole image frame. Project **A** along the two orthogonal bases $\mathbf{I_1}$ and $\mathbf{I_2}$ and compare the magnitudes of projection. If the projection along $\mathbf{I_2}$ is greater than that $\mathbf{I_1}$ then the *dominant* camera motion is due to camera *Z-rotation* (ω_Z). Otherwise, the *dominant* camera motion is due to either camera *zoom* or camara *Z-translation*.

Note that the sign of the projection in the above projection test gives the direction of camera motion. In other words, for camera *Z-rotation*, if the projection is negative, the direction of rotation is clock-wise and *vice versa.* Similarly, for camera *zoom* or camara *Z-translation*, if the projection is negative, it is either camera *zoom-out* or camera *Z-translation-away* and *vice versa.*

However, the above test can not separate the camera *zoom* from camera *Z-translation.* In order to do this, we observe that (see Eq. 7 and Eq. 9) the camera *zoom* terms in the *singular* terms of optical flow affect the magnitudes of the optical flow in non-linearly with respect to distance from the image center (x and y). This leads to the variations in the magnitudes of the optical flow vectors (note that this is a *qualitative* property). However, if we can compensate for this effect, then, we should expect the variance in the magnitudes of the compensated *zoom* vectors to be lower. But, in the case of camera *Z-translation*, since the Z is an unknown variable, the variance should be more (assuming that there is reasonable variation in the Z values). In order to compensate for effect of camera *zoom*, we use the following linear approximation[1]:

$$f_x \, tan^{-1}(\frac{x}{f_x}) \, (1 + \frac{x^2}{{f_x}^2}) \;\; \approx \;\; x. \tag{13}$$

The complete test is summarized as the following statistical test.

[1] It is important to note that this approximation has been made not just for values of x satisfying $|\frac{x}{f_x}| \ ll1$, which is valid only near the camera center, but also for those values of x satisfying $|\frac{x}{f_x}| \approx 1$, which is the case for many normal to wide-angled shooting. For this, we approx imate $tan^{-1}(\frac{x}{f_x})$ as $(\frac{x}{f_x} - (\frac{1}{3})(\frac{x}{f_x})^3)$, which leads to $f_x tan^{-1}(\frac{x}{f_x})(1 + \frac{x^2}{{f_x}^2})$ being approximated as $x(1 + (\frac{2}{3})(\frac{x}{f_x})^2 - (\frac{1}{3})(\frac{x}{f_x})^4)$. Further observation confirms that the effect of the quadratic and fourth-order terms in the bracket are only of the order of the constant term 1 at least until the values of x satisfy $|\frac{x}{f_x}| \approx 1$. Hence, from the point of view of reducing the *singular variance* after compensation of camera zoom effect over the entire image frame, we feel that the above linear approximation is valid. However, this linear approximation may not be valid for videos taken with extremely wide field of view.

Singular variance test

Divide the x-component of $\mathbf{OF}(\mathbf{x}, \mathbf{y})$ by x and the y-component by y (to compensate for the effect due to camera *zoom*) and then compute the magnitude of the observed optical flow vectors. Compute the mean and variance of the magnitudes. If the variance is more than a threshold $Tvar_{singular}$ (most likely caused by the variation in the Z parameter involved in the description of *Z-translation*), then the observed *dominant* flow is due to camera *Z-translation.* Otherwise (it is more likely that there is no Z parameter involved in the description of the observed *singular* flow) the observed dominant flow is due to camera *zoom.*

2.2 Subclassification of Non-Singular Flow

As can be seen from Eq. 8 and Eq. 10, it is clear that the non-singular flow terms do not depend on x and y coordinates at all, and that the most important parameters governing the *non-singular* flow are (i) camera rotations: *pan* parameter ω_Y and *tilt* parameter ω_X, and (ii) camera translations: *horizontal* translation T_X and *vertical* translation T_Y. So, if a *nonsingular* has been observed, it is likely that the *dominant* camera motion has been due to one or all of these 4 parameters only. A closer observation will show that only the term(s) involving T_X and/or T_Y contain(s) Z in the denominator.

Now, in order to sub-classify the *non-singular* flow, as in the preceding subsection, we compute the 2D affine motion model consiting of 6 parameters. Then, since the observed dominant flow is *non-singular*, we expect the norm of $\mathbf{A}$ to be lower and the norm of $\mathbf{b}$ to be more than a threshold $Th_{non-singular}$. The *magnitude* and *sign* of $\mathbf{b}$ then indicate whether the camera motion is *left* or *right* or up or *down* or any of the 4 possible combinations of these.

Remark 2.2 *Note that, for the purpose of sub-classification of* dominant non-singular *flow, we could have computed only* $\mathbf{b}$ *above instead of 6 parameters (both* $\mathbf{A}$ *and* $\mathbf{b}$*). However with the 6 parameters, we can also identify other possible* secondary singular *motion content, which are observed in some rare* pan-zoom *combinations in a video segment, by analysing parameters in* $\mathbf{A}$ *as described in Section 2.1.*

Assuming the scene contains reasonable variation in the values of Z, we design a classification method for further identification of whether the dominant camera motion is *rotation* (*pan* or *tilt* or their combination) or *translation* (*horizontal* or *vertical* translations or their combination) by the following statistical test.

Non-singular variance test

Compute the magnitude of the observed optical flow vectors. Compute the mean and variance of the magnitudes. If the variance is more than a threshold $Tvar_{non-singular}$ (most likely caused by the variation in the Z parameter involved in the description of *non-singular* flow), then the observed dominant flow is due to camera *translation(s)*. Otherwise (it is more likely that there is no Z parameter involved in the description of the observed *non-singular* flow) the observed dominant flow is due to camera *rotations(s)*.

3 Motion Measurement

In the previous section, we described our approach to interpret the *dominant* camera motion under the assumption that optical flow corresponding to camera motion has been already measured. However, accurate measurement of motion (which in our case is optical flow on the image coordinates) is well-known to be very difficult [18, 19]. One of the most common approaches is to work under the *brightness constancy* assumption which gives us the following optical flow equation [20, 21, 22].

$$I_x(x,y)\,u(x,y) + I_y(x,y)\,v(x,y) + I_t(x,y) = 0 \tag{14}$$

where $I_x(x,y)$ is the x-gradient, $I_y(x,y)$ is the y-gradient and $I_t(x,y)$ is the t-gradient (temporal gradient) of the image intensity. These image gradients are computed using suitable discrete difference schemes [23]. In the following subsection, we describe our method to compute the optical flow using the above optical flow equation.

3.1 Computation of Optical Flow

It is clear that, at each point $\mathbf{p} = (\mathbf{i}, \mathbf{j})^{\mathbf{t}}$ on the image coordinates, Eq. 14 will give only one equation whereas there are two unknowns to compute for $\mathbf{OF}(\mathbf{i},\mathbf{j}) = (\mathbf{u}(\mathbf{i},\mathbf{j}), \mathbf{v}(\mathbf{i},\mathbf{j}))^{\mathbf{t}}$. This makes the measurement of optical flow an *under-constrained* problem. In other words, only that component of the optical flow, which is along the image gradient direction $(I_x(i,j), I_y(i,j))^t$ at pixel coordinates (i,j) can be directly measured. (This problem is also known as the *aperture* problem in literature and the measured flow is refered to as *normal flow* in literature since it is normal to the local edges [20, 15, 19].) In order to overcome this problem, researchers have used various kinds of *smoothness* assumptions also [20, 22, 19]. In our approach, we assume the *local-smoothness* of optical flow over a local region (we use 7×7 window centered on the image coordinates (i, j)) and solve for the optical flow using linear least square error technique [23, 24]. In what follows, we describe our method of computation of optical flow at a pixel (i, j).

Let the least square error estimate for optical flow at pixel (i, j) be given by $\mathbf{\hat{OF}(i, j)} = (\mathbf{\hat{u}(i, j)}, \mathbf{\hat{v}(i, j)})^{\mathbf{t}}$. Assuming this estimate to be a common (hence smooth) solution to a local region R (we use 7×7 neighborhood) centered around (i, j), we can list a set of over-determined optical flow constraint equations as

$$\mathbf{P}\ \mathbf{\hat{OF}(i, j)} \;=\; \mathbf{Q} \tag{15}$$

where

$$\mathbf{P} \;=\; \begin{pmatrix} I_x^1 & I_y^1 \\ I_x^2 & I_y^2 \\ & . \\ & . \\ & . \\ I_x^N & I_y^N \end{pmatrix}, \quad \mathbf{\hat{OF}(i, j)} \;=\; \begin{pmatrix} \hat{u}(i, j) \\ \hat{v}(i, j) \end{pmatrix}, \quad \textbf{and}, \quad \mathbf{Q} \;=\; \begin{pmatrix} -I_t^1 \\ -I_t^2 \\ . \\ . \\ . \\ -I_t^N \end{pmatrix}$$

where, N is the number of pixels in the local region R. The least square error solution is given by

$$\mathbf{\hat{OF}(i, j)} = \left(\mathbf{P^t\, P}\right)^{\mathbf{-1}} \mathbf{P^t\, Q}. \tag{16}$$

Remark 3.1 *The above method of computation of optical flow inherently assumes two things. The first one is the* local-smoothness *of optical flow. This is quite justifiable, in most area on the image, in the case of computation of optical flow due to* global *camera motion. The second assumption is that, along with the first assumption, we are hoping to overcome the so called* aperture *problem by listing many (more precisely, an over-determined set) optical flow constraint equations in a local region (we use* 7×7 *window centered on the pixel where least square error optical flow is being computed). This assumption is basically equivalent to* expecting *reasonable texture in the local area which means we expect that the gradient direction varies sufficiently the local area for the over-determined set to capture the complete optical flow (as against only the normal flow) through the observation of optical flow along* linearly independent *gradient directions. It is important to note that just two such linearly independent directions (which are also maximum possible, given that the image is 2D) suffice. If this assumption fails, the computed optical flow is only the normal flow.*

It is important to note that Eq. 14 is valid only in the limit. Hence, in the computation of optical flow, the frames are always adjacent ones (even for frame rates of 30 frames per second). This aspect of optical flow has both merits and demerits. The merit is, it enables us to derive at least one valid constraint for its measurement (Eq. 14). The demerit, which is very relevant to our purpose of *dominant* motion interpretation, is that the measured values of components of optical flow are quite small and noisy between adjacent frames, and hence may lead to serious errors in the computation of any motion model parameters to even *qualitatively* describe the optical flow. As described in the previous section, we need to observe a robust optical flow in order to (i) reliably detect whether a *singular* flow field has been observed or a *non-singular* flow field has been observed and (ii) reliably estimate the 6 parameters of the 2D affine motion model to *qualitatively* capture the nature of camera work using which further sub-classification of camera work is done. In order to increase the robustness in the above two steps, we compute *Optical Flow Streams* (OFS) as described in the following subsection.

3.2 Computation of OFS

Optical flow streams are just streams of optical flow linked through a specified number of temporal frames. More precisely, we define

$$\mathbf{OFS(i,j)} = \sum_{\mathbf{k=1}}^{\mathbf{M-1}} \mathbf{OF(i_k,j_k)} \tag{17}$$

where

$$\begin{pmatrix} i_k \\ \\ j_k \end{pmatrix} = Round\left\{ \begin{pmatrix} i \\ \\ j \end{pmatrix} + \sum_{p=1}^{k-1} \mathbf{OF(i_p,j_p)} \right\}.$$

Here, $\mathbf{OFS(i,j)}$ is the optical flow stream at a pixel (i,j) in current frame indexed by $k = 1$, which is computed by linking the optical flow through M frames. In the above equation, a common factor Δt (the time between adjacent frames) of the term $\sum_{p=1}^{k-1} \mathbf{OF(i_p,j_p)}$ has been omitted because $\Delta t = 1$ since, obviously, the optical flow between adjacent frames is computed in units of pixels per Δt. Note that $\mathbf{OFS(i,j)}$ is a vector sum and that the $Round\{\}$ operation described rounds each component of the 2D vector to the nearest integer, so that it identifies an image pixel.

We wish to stress that the effect of computation of OFS is *robustness* in extracting the *qualitative* properties of the optical flow. This robustness is achieved through temporal integration (and hence smoothing) which results not only in (i) overcoming the noise that affects the computation of image gradients by discrete difference scheme and thereby the computation of the optical flow field but also in (ii) the observation of significant magnitudes of the optical flow (it is important to note that optical flow between adjacent frames is quite small and noisy to be relied upon) from which the *qualitative* properties (like (i) detection of the singularity, (ii) extraction of 2D affine motion model parameters for further *qualitative* sub-classification of both *singular* and *non-singular* flow) can be robustly extracted. This is evidenced from the various experimental results reported in Section 5.

Remark 3.2 *It should be noted that, in the computation of OFS, we inherently assume that the* dominant *motion in all of the M frames is* qualitatively *unchanging. While this assumption does not hold across* camera breaks *[25, 26, 27], it holds for a major portion video segment thus advocating its use in view of its robustness for motion classification purposes. Moreover, by choosing an appropriate value for M, it is possible to maintain both robustness and accuracy of* qualitative *motion interpretation. In all our experimental results reported in Section 5, we use a value of* 10 *for M which amounts to temporal integration for $\frac{1}{3}^{rd}$ of a second for video data taken at* 30 *frames per second.*

4 The complete algorithm

Based on our approach described in detail in the two previous sections, we now summarize our automatic video motion annotation algorithm (see Figure 4). As shown in Figure 4, we compute the OFS projections on each frame of the video segment and then analyse each frame independently for further classification. For the computation of the OFS projection image, we consider only the magnitudes of OFS vectors over the image coordinates, compare with a binarization threshold and output a binary image containing gray values 0 and 255 only (255 is used if the magnitude exceeds the threshold). For further classification, first, we detect whether camera motion has occured or not using the binarized OFS projection image. If the percentage of pixels set to gray value 255 exceeds a threshod $Th_{camera-work}$, we classify it as camera motion. Otherwise, if the percentage of pixels is less than $Th_{object-motion}$, we classify it as object motion which is further checked and classified as no motion if the percentage of pixels less than $Th_{no-motion}$ (≈ 0). Also, if the percentage of pixels exceeds $Th_{object-motion}$ but is less than $Th_{camera-work}$, we classify it as *problem class* which may have been caused due to either camera motion or many (large) object motion(s). At this stage, we analyse the frames belonging to *problem class* manually and interpret.

As shown in Figure 4, we further sub-classify the frames classified as camera work. Firstly, we detect for *singularity.* For this, we use a simple but robust method as follows.

Singularity detection

Compute the centroid $W_{centroid}$ of the white portion of the thresholded and binarized OFS projection. Compute the largest *connected* black component of the OFS

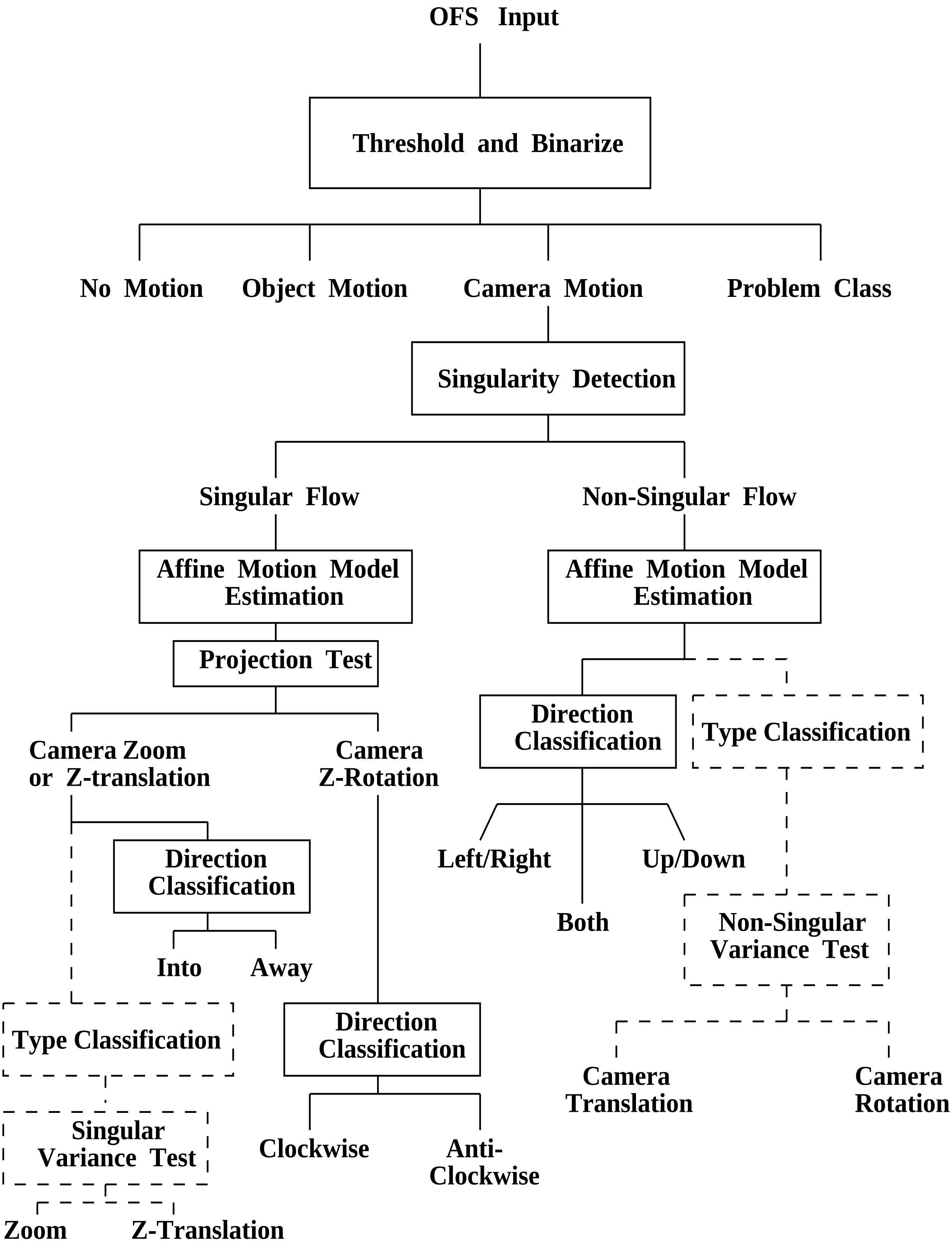

Figure 4: Complete algorithm for the proposed motion annotation scheme

projection and compute the centroid $B_{centroid}$ of it. If the distance between the two centroids is less than a threshold $Th_{category}$, then the observed optical flow is *singular*, otherwise it is *non-singular.*

Remark 4.1 *Figures 2 and the experimental results of Section 5 demonstrate the robustness of the simple singularity detection method given above. We wish to stress that we are reporting this test because it has been very effective in all our experiments reported in the next section. However, we agree that other tests, probably more robust ones, may also be designed as suitable for the purpose of singularity detection.*

After the detection of singularity, we further sub-classify as clearly shown in Figure 4. Note that we have shown in dotted lines our proposed method of separation of translation components from the rotation components, where they are applicable. Though we have described variance based tests for this purpose in Section 2, we have not had complete success in the robust separation of the two. This is due to many reasons of which the following are important:

(i) the role of the unknown variable Z in the translational components is extremely difficult to capture in general,

(i) coarse motion between adjacent frames, even in the case of 30 frames per second temporal sampling of physical phenomena, and,

(iii) the camera *zoom* modeling terms (see Eq. 4) may not be generally valid all over the image co-ordinates.

We are currently trying to overcome these problems with the possible use of domain knowledge as a powerful cue. Hence, we report our results in the next section without the dotted classification steps shown in Figure 4. We do show, however, the variances in the spatial distribution of the OFS magnitudes, to highlight the merits and demerits of the two variance based tests we have described in Section 2.

5 Experimental Results

In this section we present results on a variety of video sequences to demonstrate the effectiveness and robustness of our proposed scheme for motion annotation of video sequences. For all the presented results here, the complete algorithm described in Section 4 is used. During the computation of the optical flow and OFS, we reduce the input image size by half in each dimension so that it leads to smoothing of images and also reduces the computational burden during the computation of optical flow. For all the experiments reported in this section, we use the following thresholds:

$$\boxed{\begin{array}{ccc}
Th_{category} & = & 20 \\
Th_{singular} & = & 3 \\
Th_{non-singular} & = & 3 \\
Th_{camera-work} & = & 80 \\
Th_{object-motion} & = & 60 \\
Th_{no-motion} & = & 1
\end{array}}$$

Since these thresholds are not changed for different sequences, they highlight the robustness of the proposed scheme for video annotation by *dominant* motion interpretation.

Also, for the presentation of the motion annotation results (both automatic and manual in a single plot) for a video sequence, we use the interpretation values shown in the Table 1.

Table 1: Interpretation values used for motion annotation

Type of camera work	Interpretation value used for Automatic Annotation	Interpretation value used for for Manual Annotation
No_Motion	0	-5
Object_Motion	1	-4
Problem_Class	3	-2
Camera_Zoom_In	40	35
Camera_Zoom_Out	41	36
Camera_Move_Left	70	65
Camera_Move_Right	71	66
Camera_Move_Up	90	85
Camera_Move_Down	91	86
Camera_Move_Left_Up	110	105
Camera_Move_Left_Down	111	106
Camera_Move_Right_Up	112	107
Camera_Move_Right_Down	113	108

5.1 Case 1: *Fleet* Video Sequence

This sequence consists of 25 interpreted frames of size 150 by 150 pixels. The binarization threshold used to generate OFS magnitude signature is 1.5 (we have used smaller value here than that for other cases, where we have used a value of 2.0, because of the smaller size of the images in this case). Fig. 5 shows the result of the proposed automatic video annotation algorithm. Also shown in the figure is the manual annotation for comparison. Note that, except for the first frame which is classified as *problem_class* (this is because, after binarization, the percentage of pixels labelled 255 is slightly below the $Th_{camera-work}$), all other frames are correctly classified as camera *zoom-into.* Fig. 6 shows some frames of the sequence to visually depict the nature of camera work, and, Fig. 7 shows the typical OFS magnitude signature observed on the frames of the sequence.

Fig. 8 shows the plot of (*singular*) variances (see Section 2.1) in the spatial distribution of magnitudes of OFS computed over the image coordinates. Note that variance values are smaller which can enable the *singular* variance test of Section 2.1 to distinguish camera *zoom* from camera *Z-translation.*

5.2 Case 2: *Mall* Video Sequence

This sequence consists of 45 interpreted frames of size 512 by 256 pixels. The binarization threshold used to generate OFS magnitude signature is 2.0. Fig. 9 shows the result of the proposed automatic video annotation algorithm. Also shown in the figure is the manual annotation for comparison. Note that, except for one frame which is classified as *problem_class* (this is also because, after binarization, the percentage of pixels labelled 255 is slightly below the $Th_{camera-work}$), all other frames are correctly classified as camera *pan-left.* Fig. 10 shows some frames of the sequence to visually depict the nature of camera work, and, Fig. 11 shows the typical OFS magnitude signature observed on the frames of the sequence.

Fig. 12 shows the plot of (*non-singular*) variances (see Section 2.2) in the spatial distribution of magnitudes of OFS computed over the image coordinates. Note that variance values are smaller which can enable the *non-singular* variance test of Section 2.2 to distinguish camera *pan* from camera translation.

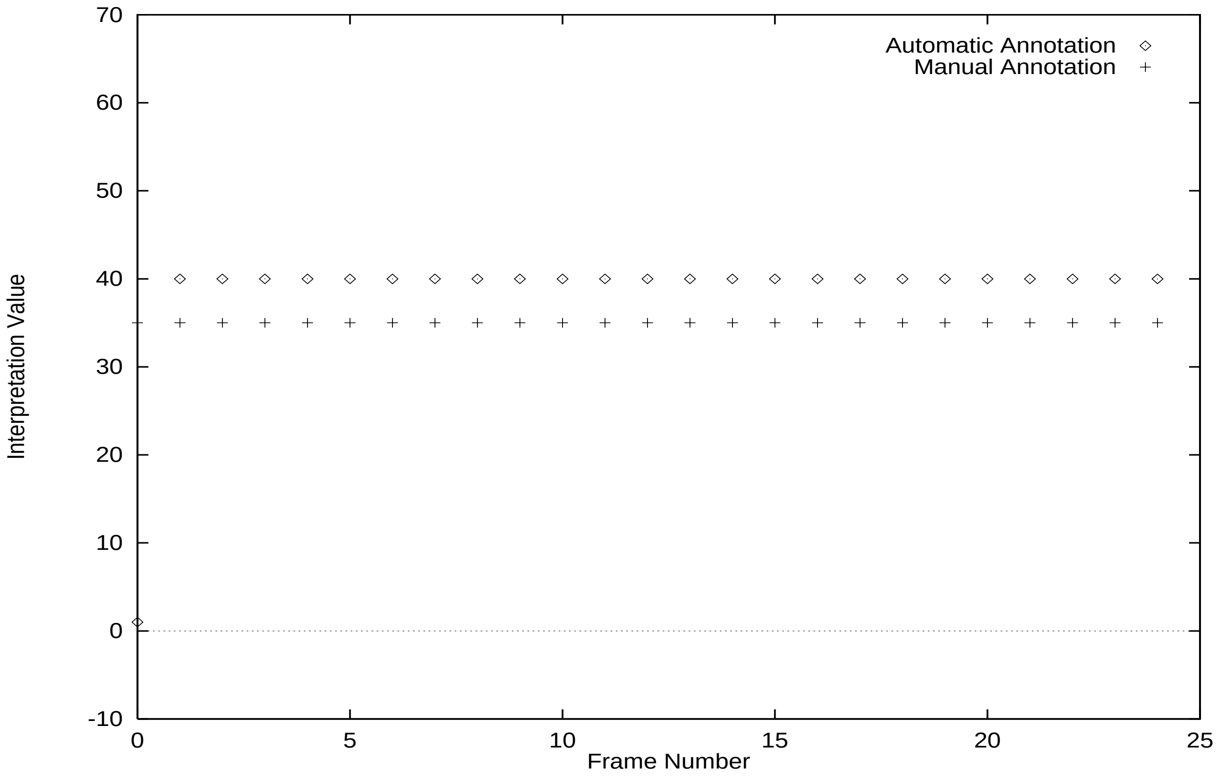

Figure 5: Motion annotation of *Fleet* sequence

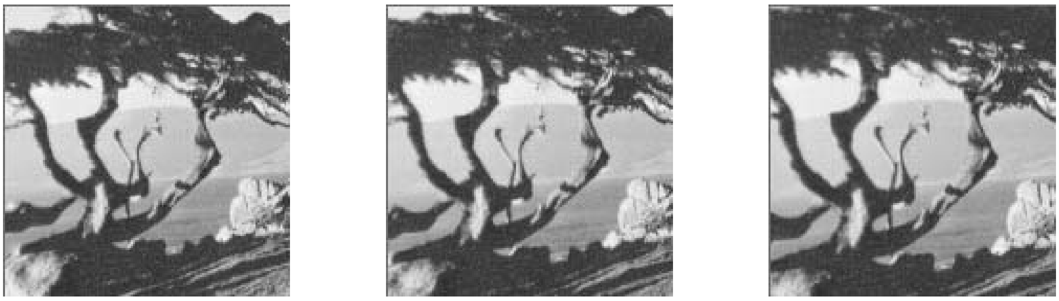

Figure 6: *Fleet* sequence: Frames 10, 15 and 20 from left to right (camera *zoom-in*)

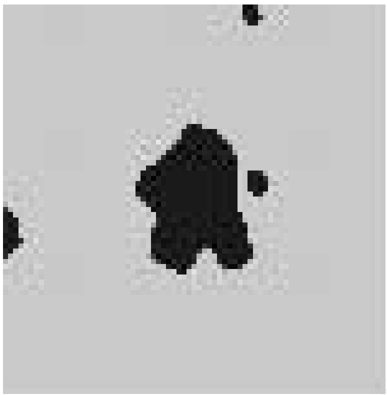

Figure 7: *Fleet* sequence: OFS magnitude signature (thresholded and binarized) measured on Frame 10

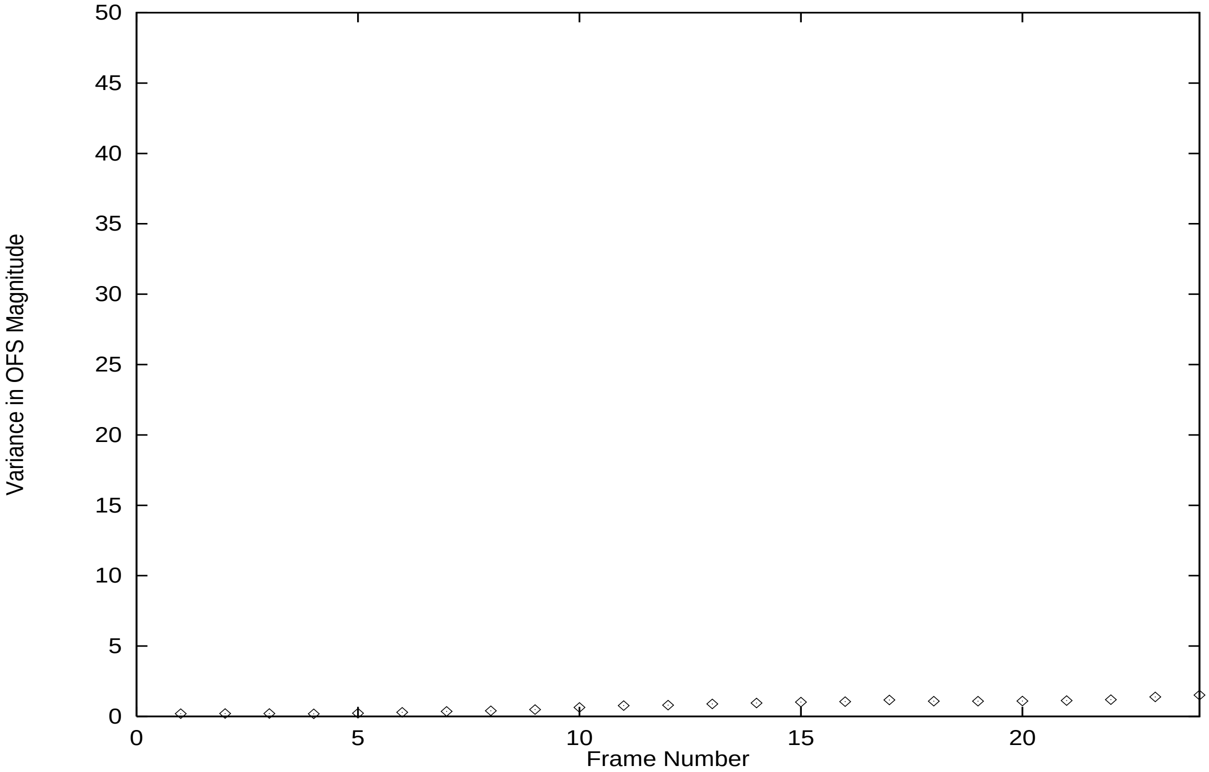

Figure 8: *Fleet* sequence: *Singular* variances in the distribution of OFS magnitudes in video frames

5.3 Case 3: *Flower Garden* Video Sequence

This sequence consists of 80 interpreted frames of size 360 by 243 pixels. The binarization threshold used to generate OFS magnitude signature is 2.0. Fig. 13 shows the result of the proposed automatic video annotation algorithm. Also shown in the figure is the manual annotation for comparison. Note that, except for two frames which are classified as *problem_class* (this is again because, after binarization, the percentage of pixels labelled 255 is slightly below the $Th_{camera-work}$), the *dominant* motion theme is correctly extracted. The first few frames are classified as camera move-right instead of camera move-right-up because the horizontal motion is much larger than the vertical motion. Noth that all other frames are correctly classified as camera move-right-up. Fig. 14 shows some frames of the sequence to visually depict the nature of camera work, and, Fig. 15 shows the typical OFS magnitude signature observed on the frames of the sequence.

Fig. 16 shows the plot of (*non-singular*) variances (see Section 2.2) in the spatial distribution of magnitudes of OFS computed over the image coordinates. Note that variance values are larger in this case which can enable the *non-singular* variance test of Section 2.2 to distinguish camera translation from camera rotation.

5.4 Case 4: *Table Tennis* Video Sequence

This sequence consists of 280 interpreted frames of size 360 by 240 pixels. The binarization threshold used to generate OFS magnitude signature is 2.0. Fig. 17 shows the result of the proposed automatic video annotation algorithm. Also shown in the figure is the manual annotation for comparison. It should be noted that, except for 5 frames which are classified as *problem_class* (the reason is again similar to that of earlier cases), the automatic annotation performance is quite good. Also, note the wrong interpretation of a few frames during the transition in camera work nature. This is because we compute OFS from the optical flow of 10 successive frames. However, the *dominant motion theme* is accurately represented in different segments of the video - first there is only object motion, then there is camera *zoom-out* followed by only object motion after which there is camera move-right. Fig. 18 shows some frames of the sequence to visually depict the camera work nature at the four different segments of the video data, and, Fig. 19 shows the typical OFS magnitude signatures observed on the frames of the sequence.

Fig. 20 shows the plot of variances (both *singular* and *non-singular*) in the spatial distribution of magnitudes of OFS computed over the image coordinates. The plot corresponding to camera *zoom-out* segment is for *singular variance* (see Section 2.1) where as the plot corresponding to camera move-right segment is for *non-singular variance* (see Section 2.2).

5.5 Case 5: *Foreman* Video Sequence

This sequence consists of 280 interpreted frames of size 176 by 144 pixels. The binarization threshold used to generate OFS magnitude signature is 2.0. Fig. 21 shows the result of the proposed automatic video annotation algorithm. Also shown in the figure is the manual annotation for comparison. In this case also, except for a few frames which are either misclassified or classifed as *problem_class* when there is inconsistent or abrupt camera motion or during transition in camera work nature, the performance of the proposed scheme is quite good in the sense that the *dominant motion theme* is accurately represented in different segments of the video - first there is mainly object motion, then there is camera move-right followed by camera move-right-down after which there is no motion of the camera or object.

Fig. 22 shows some frames of the sequence to visually depict the nature of camera work in the four different segments of the video data, and, Fig. 23 shows the typical OFS magnitude signatures observed on the frames of the sequence.

Fig. 24 shows the plot of (*non-singular*) variances in the spatial distribution of magnitudes of OFS computed over the image coordinates (only in the segments where there is camera move-right or camera move-right-down). Note that variance values are larger eventhough the camera motion is *pan* and/or *tilt* which makes it difficult for the *non-singular* variance test of Section 2.2 to distinguish camera rotation from translation.

6 Conclusion

We have presented a new approach to automatic annotation of video sequences by *dominant* motion interpretation. Unlike others, we separate the optical flow into two categories - *singular* and *non-singular* - which as we have shown is a more natural way of classification for the purpose of *dominant* motion interpretation. We have shown that identification of patterns created by such natural categorites which can be observed from the *measured* optical flow can help focus the interpretation of *dominant* motion in video segments. This amounts to *top-down* interpretation step suitably applied with *bottom-up* computations (optical flow is computed by local optimization of the optical flow constraint), in order to simplify the extraction of the *dominant* motion theme. For robust detection of such natural patterns, we have proposed the computation of so called *optical flow streams* (OFS) by linking the optical flow through M successive frames. We have demonstrated the use of OFS and their signatures when projected on to the images, not only in (*i*) the detection of camera motion as against only object motion or no motion and (*ii*) robust detection of the *qualitative* property of singularity, but also in robust sub-classification of both the *singular* and *non-singular* optical flow patterns. During each step of proposed scheme of motion annotation, we have used only the *qualitative* properties of measured optical flow (and its derivative OFS), namely the *magnitude*, *sign* and *distribution* on the image plane, of the OFS magnitudes, and hence, even in the cases where only normal flow can be computed over some portions of the images, we get reliable interpretation of the *dominant* motion theme, even though most accurate estimation of the camera motion parameters may be difficult in such cases. This is evidenced from the robustness and accuracy of automatic motion annotations obtained in the various experimental results reported in this paper.

We have proposed two statistical tests in Section 2 for the possible separation of camera rotations from the camera translations in each of the two natural categories. However, as mentioned in Section 4 and discussed with each case of the results in Section 5, we have not had complete success in the separation of camera translation from camera rotation as applicable during the motion interpretation. This is well-known to be a very difficult problem in general since the depth values Z, which affect the translational components of the observed optical flow are unknown in general [15, 12]. We are currently trying to explore the possible use of domain knowledge in order to achieve robust separation of camera translations from camera rotations.

Also, domain knowledge, along with the OFS signature, can help us accurately interpret the *dominant* motion content in video segment in many cases. For example, consider an outdoor scenario where the camera is mounted on a car undergoing purely *Z-translation*. In this case, since the sky is observed in the field of view at the top of image frames and since the distance Z to sky is relatively infinite (see Remark 2.1), it would be impossible to observe *global* motion on image plane in the first place, after which correct sub-classification may indeed be possible. This is illustrated in Fig. 26 which shows the OFS magnitude signature measured on a frame of the video sequence shown in Fig. 25. In this case, since sky takes a large portion at the top of the images and does not produce any significant motion on the image plane from camera *Z-translation*, it would be difficult to classify whether it is (large) object motion or camera motion. It is important to note that, if this is classified into *singular* category, further classification may indeed be quite easy using the scheme described in Section 2.1. In this context, domain knowledge, say for example the knowledge that the segment belongs to outdoor scenario, would help one to ignore the OFS magnitude signature at the top most portion of the images using a suitable scheme and thus be able to classify measured flow first as camera work and then as *singular* category after which the sub-classification becomes easier.

The above example is just to illustrate the merit of using domain knowledge in conjunction with OFS signatures which can be robustly extracted. It should be noted that the use of domain knowledge also helps in many cases of sub-classification of the *object* motion. As part of our ongoing *Videobook* project, we are currently doing research on such domain specific extraction of motion indices, to annotate the video frames with *object* motion content also, in addition to the presently reported scheme, utilizing the robustly observable OFS signatures from the motion of objects.

Acknowledgement

We thankfully acknowledge the support of Sino Software Research Center of Hong Kong University of Science & Technology for carrying out this research under grant SSRC94/95.EG08.

References

[1] S. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *Multimedia*, vol. 1, no. 2, pp. 356–365, 1994.

[2] A. Bimbo, E. Vicario, and D. Zingoni, "Sequence retrieval by contents through spatio temporal indexing," in *Proc. IEEE Symposium on Visual Languages*, pp. 88–92, 1993.

[3] R. Jain and A. Hampapur, "Metadata in video databases," *ACM SIGMOD Record*, vol. 23, no. 4, 1994.

[4] M. J. Swain and D. H. Ballard, "Color Indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[5] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QBIC project: Query images by content using color, texture and shape," in *SPIE Proc. Storage and retrieval for image and video databases*, vol. 1908, pp. 173–186, 1993.

[6] A. Pentland, R. Picard, and S. Scaroff, "Photobook: Tools for content-based manipulation of image databases," in *SPIE Proc. Storage and retrieval for image and video*

databases II, vol. 2185, pp. 34–46, 1994. Longer version available as MIT Media Lab Perceptual Computing Technical Report No.255, Nov. 1993.

[7] C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image, Vision and Computing*, vol. 13, no. 2, 1995.

[8] Y. Tse and R. Baker, "Global zoom/pan estimation and compensation for video compression," in *ICASSP*, vol. 4, pp. 2725–2728, 1991.

[9] S. Wu and J. Kittler, "A differential method for simultaneous estimation of motion, change of scale and translation," *Signal Processing: Image Communiication*, vol. 2, pp. 69–80, 1990.

[10] M. Hoetter, "Differential estimation of global parameters pan and zoom," *Signal Processing*, pp. 249–265, 1989.

[11] J. Park *et al*, "A differential method for simultaneous estimation of motion, change of scale and translation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, 1994.

[12] M. Srinivasan, S. Venkatesh, and R. Hosie, "Qualitative estimation of camera motion parameters from video sequences," *Pattern Recognition Special Issue (to appear)*, 1996.

[13] K. Rangarajan and M. Shah, "Interpretation of motion trajectories using focus of expansion," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 14, no. 12, pp. 1205–10, 1992.

[14] A. Akutsu *et al.*, "Video indexing using motion vectors," in *SPIE Proc. Visual Communication and Image Processing'92*, vol. 1818, pp. 522–530, 1992.

[15] C. Fermuller, "Global 3D motion estimation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 415–421, 1993.

[16] H. Longuet-Prazdny and K. Prazdny, "The interpretation of a moving retinal image," *Proc. R. Soc. London Ser. B*, vol. 208, pp. 385–397, 1980.

[17] A. Verri, F. Girosi, and V. Torre, "Mathematical Properties of Two-dimensional Motion Field: from Singular points to Motion Parameters," *J. of Opt. Soc. of Am. Series A*, vol. 6, pp. 698–712, 1989.

[18] J. Aggarwal and N. Nandhakumar, "On the computation of motion from sequences of images - a review," *Proceedings of the IEEE*, vol. 76, pp. 917–935, 1988.

[19] J. Barron *et al*, "Performance of optical flow techniques," *International Journal on Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[20] B. Horn and B. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[21] B. Horn and B. Schunck, "Determining Optical Flow: A Retrospective," *Artificial Intelligence*, vol. 59, pp. 81–87, 1993.

[22] A. Verri, F. Girosi, and V. Torre, "Differential Techniques for Optical Flow," *J. of Opt. Soc. of Am. Series A*, vol. 7, pp. 912–922, 1990.

[23] G. Sudhir, S. Banerjee, R. Bahl, and K. Biswas, “A Cooperative Integration of Stereopsis and Optic Flow Computation,” *J. of Opt. Soc. of Am. Series A*, vol. 12, p. 2564, 1995.

[24] E. Chalom and V. M. B. Jr., “Segmentation of Frames in Video using Motion and Other Attributes,” techincal report, M.I.T. Media Laboratory, 1995.

[25] J. C. Lee and M. Ip, “A robust approach for camera break detection in color video sequence,” in *Proc. IAPR Workshop on Machine Vision Application (MVA'94)*, 1994.

[26] H. J. Zhang, A. Kankanhalli, and S. Smoliar, “Automatic partitioning of full-motion video,” *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, 1993.

[27] W. Xiong, J. C. Lee, and M. Ip, “Net comparison: A fast and effective method for classifying image sequence,” in *Proc. IS&T/SPIE Symposium on Storage and Retrieval for Image and Video Databases*, 1995.

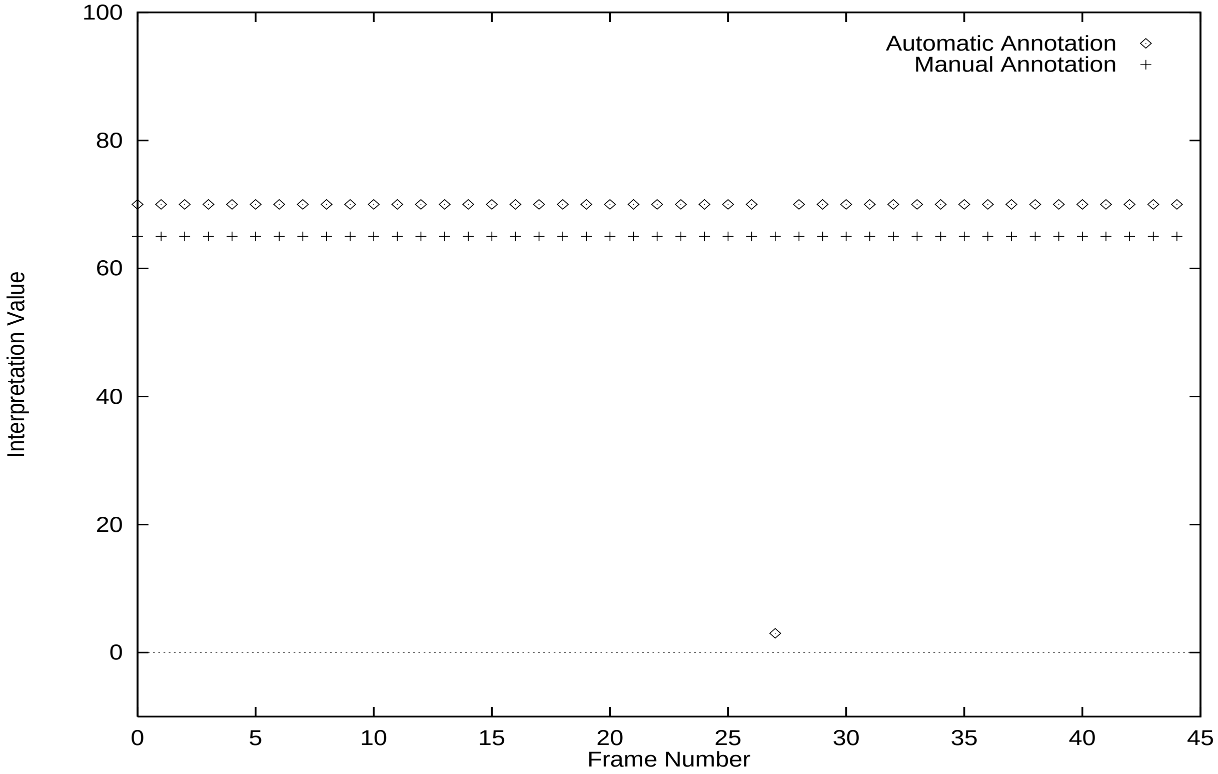

Figure 9: Motion annotation of *Mall* sequence

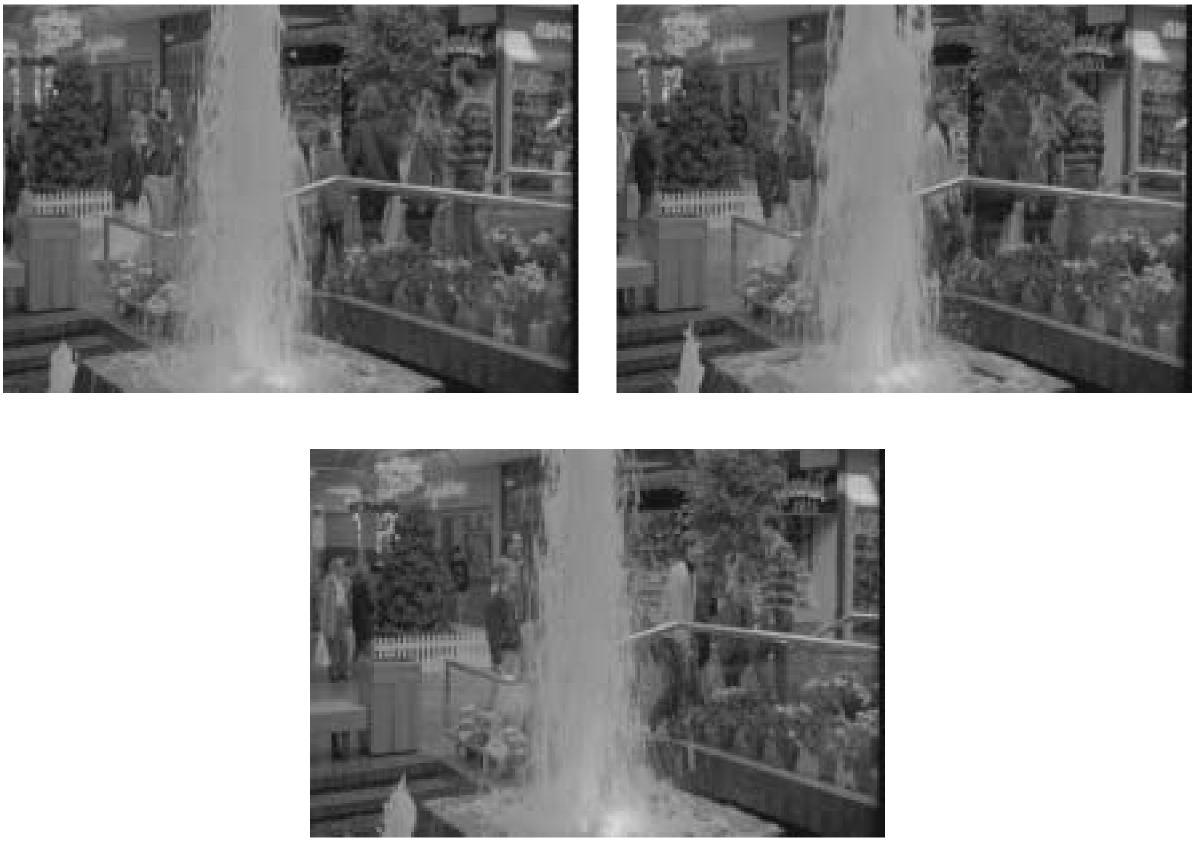

Figure 10: *Mall* sequence: (a) top-left: Frame 1, (b) top-right: Frame 10 and (c) bottom: Frame 20 (camera *pan-left*)

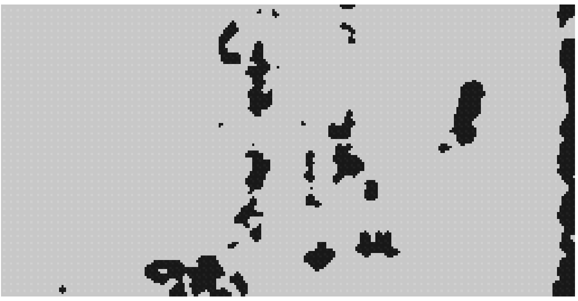

Figure 11: *Mall* sequence: OFS magnitude signature (thresholded and binarized) measured on Frame 20

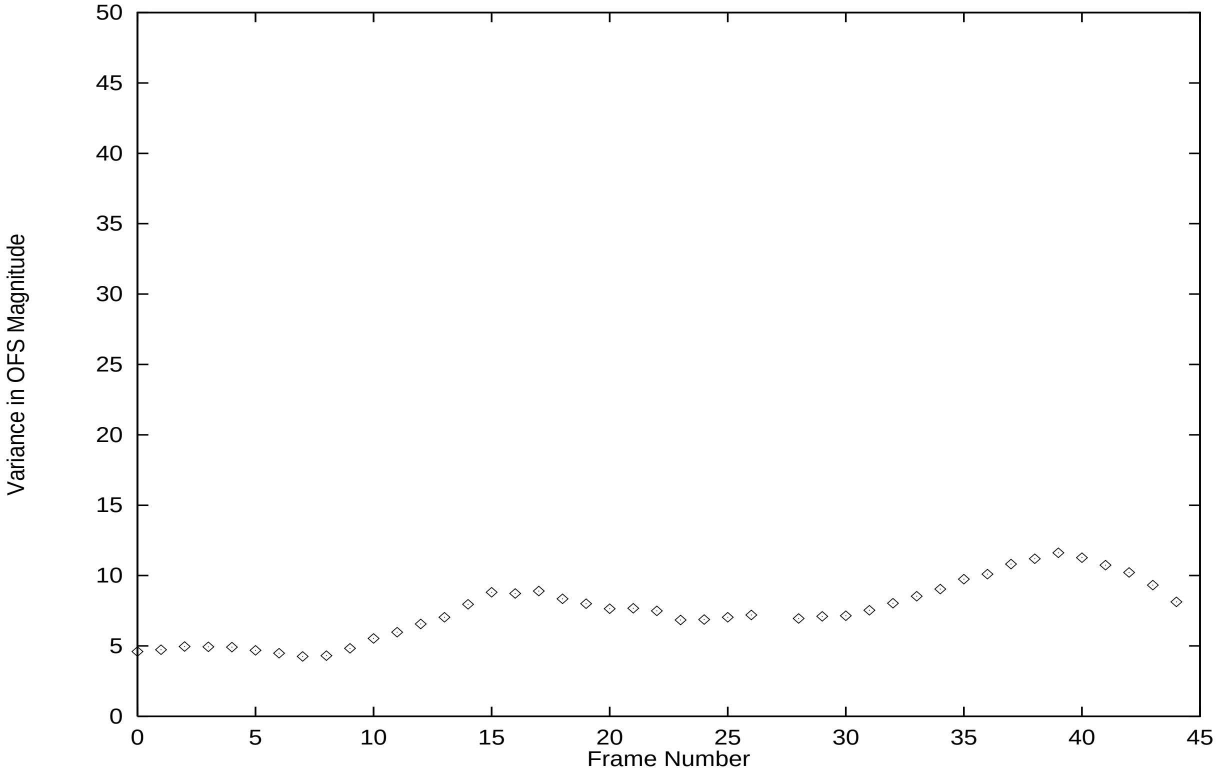

Figure 12: *Mall* sequence: *Non-Singular* variances in the distribution of OFS magnitudes in video frames

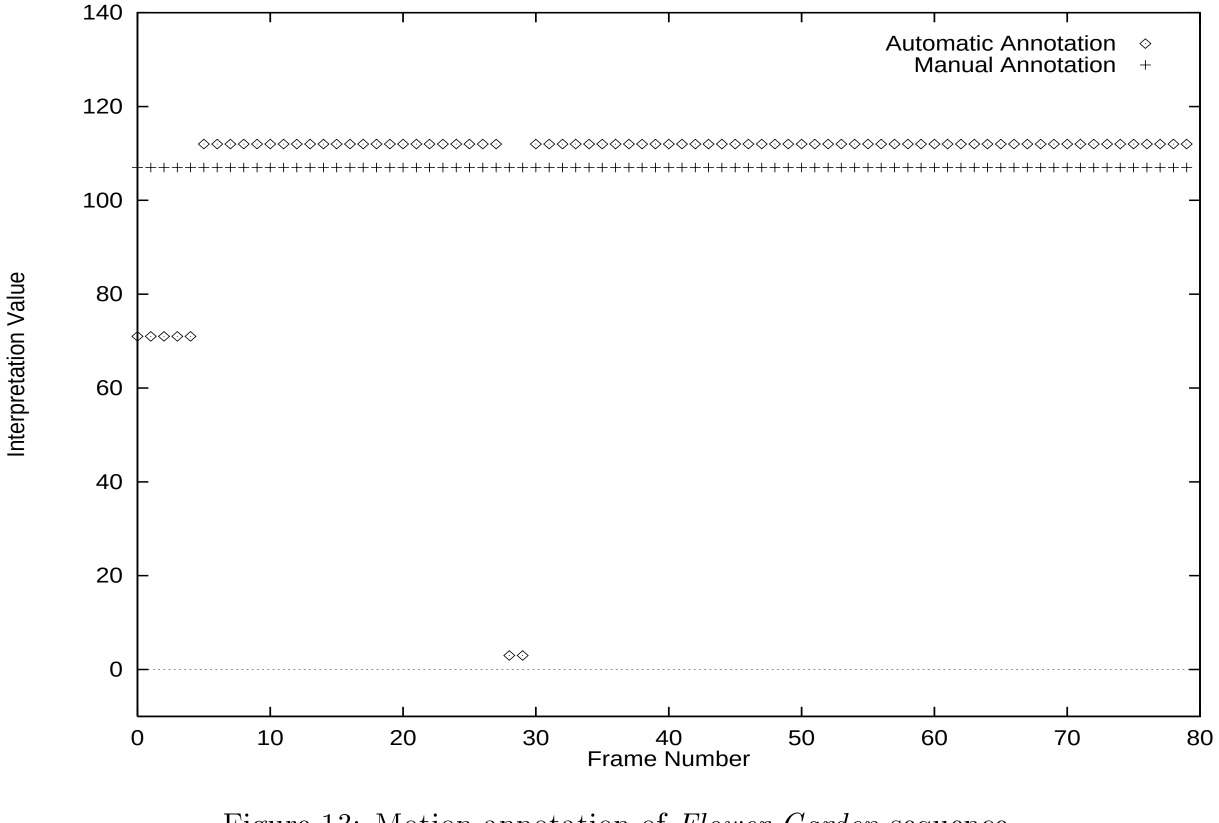

Figure 13: Motion annotation of *Flower Garden* sequence

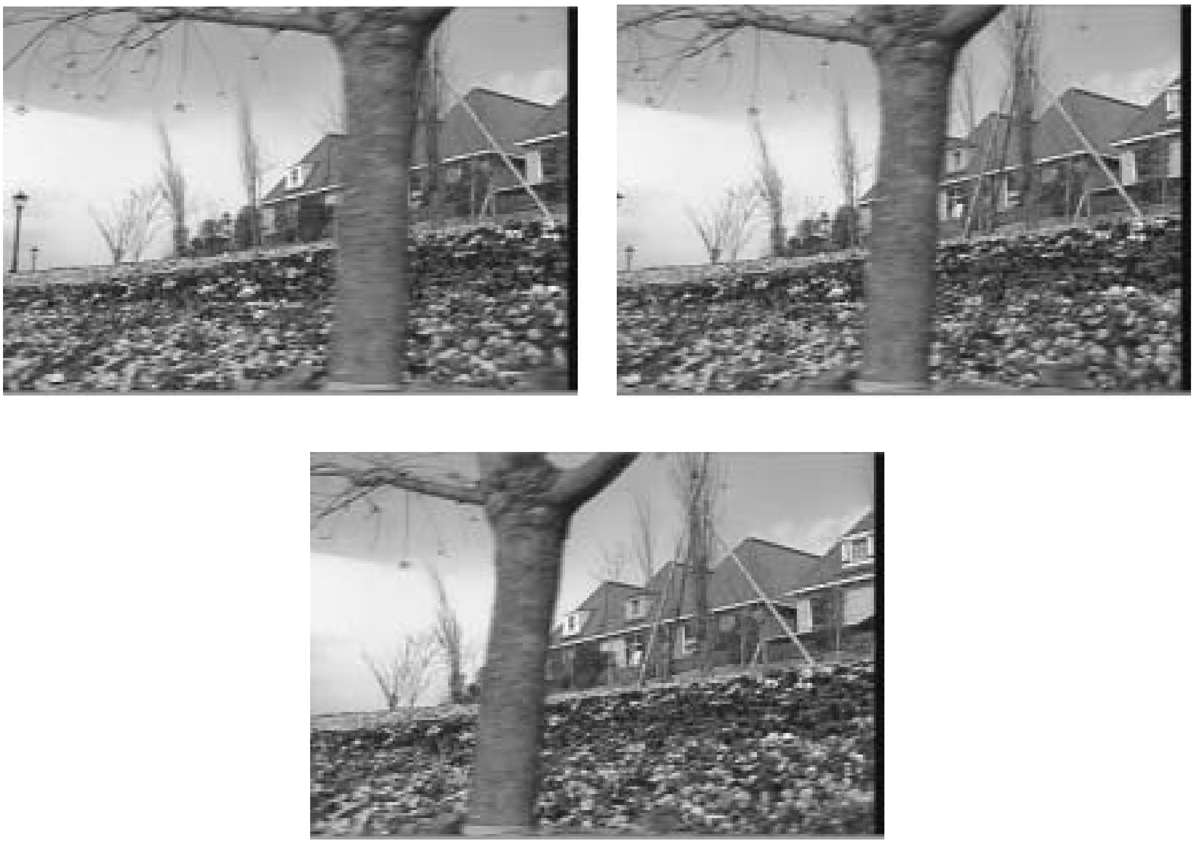

Figure 14: *Flower Garden* sequence: (a) top-left: Frame 1, (b) top-right: Frame 10 and (c) bottom: Frame 20 (camera translation)

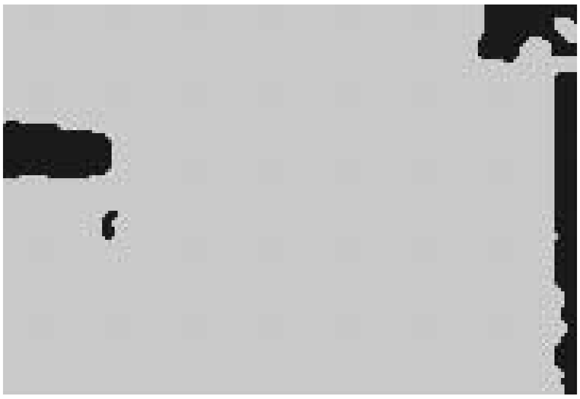

Figure 15: *Flower Garden* sequence: OFS magnitude signature (thresholded and binarized) measured on Frame 10

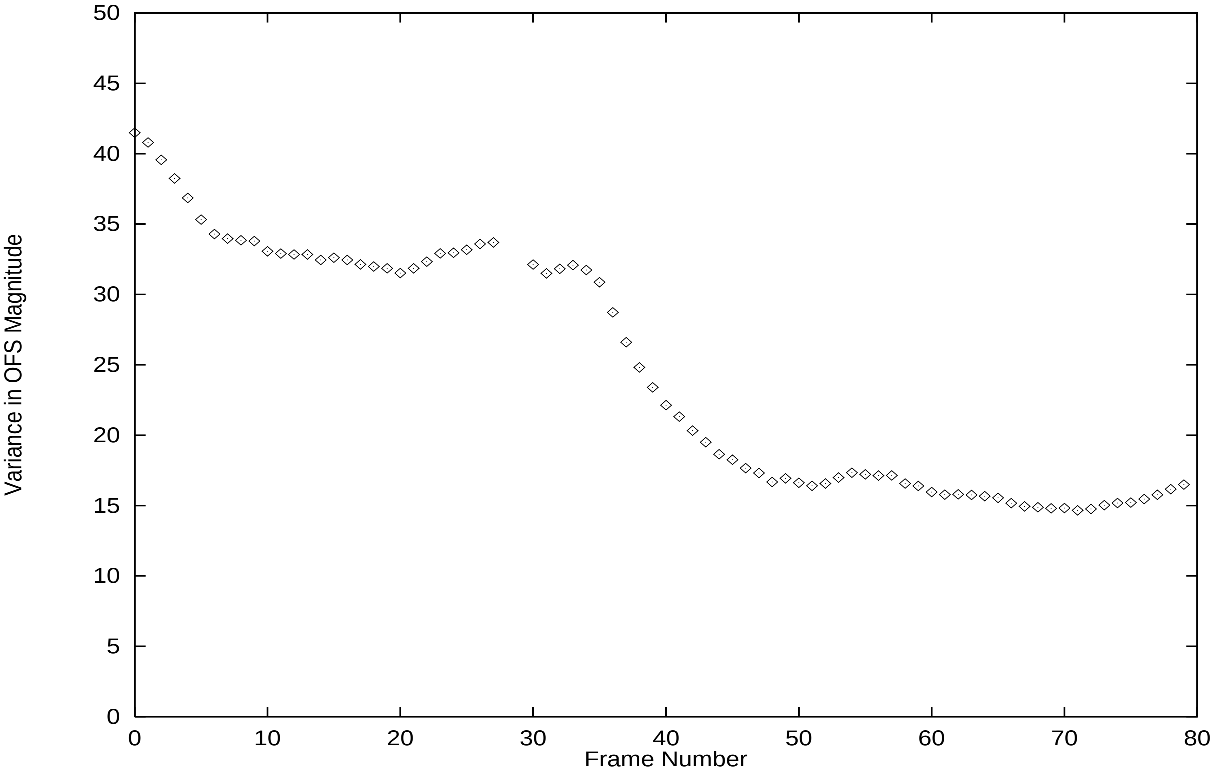

Figure 16: *Flower Garden* sequence: *Non-Singular* variances in the distribution of OFS magnitudes in video frames

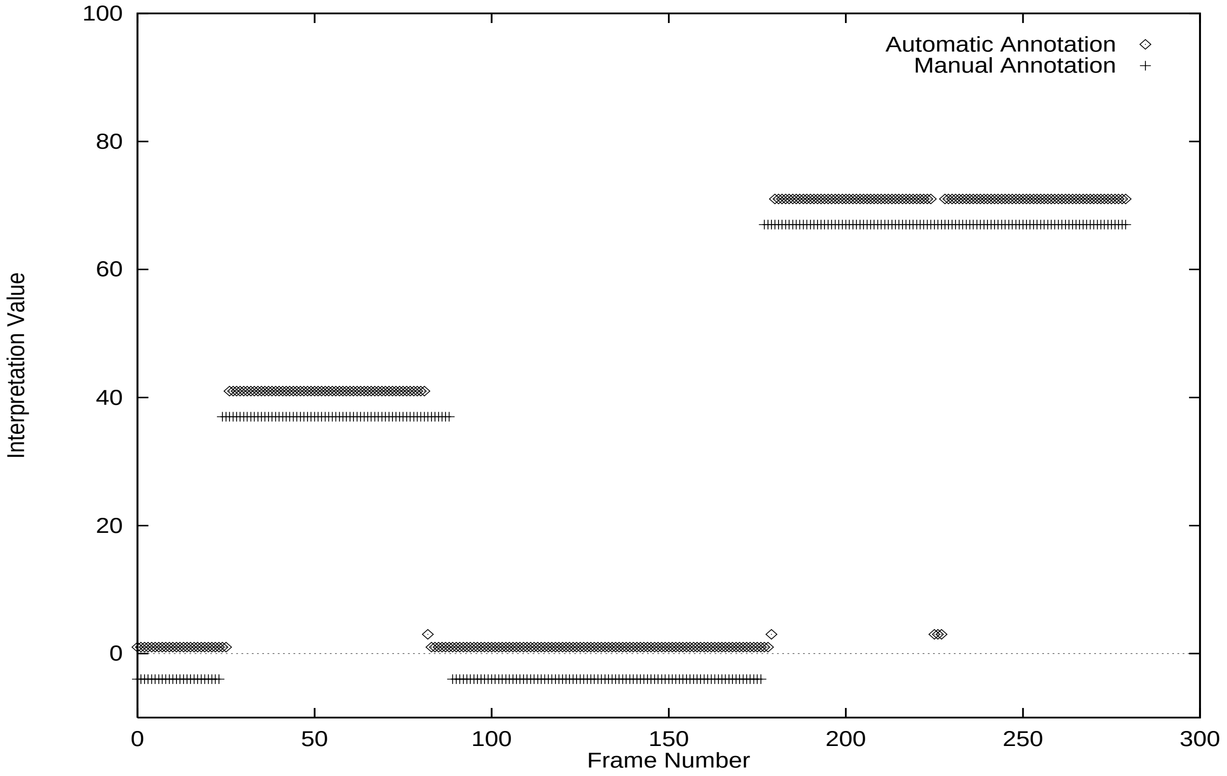

Figure 17: Motion annotation of *Table Tennis* sequence

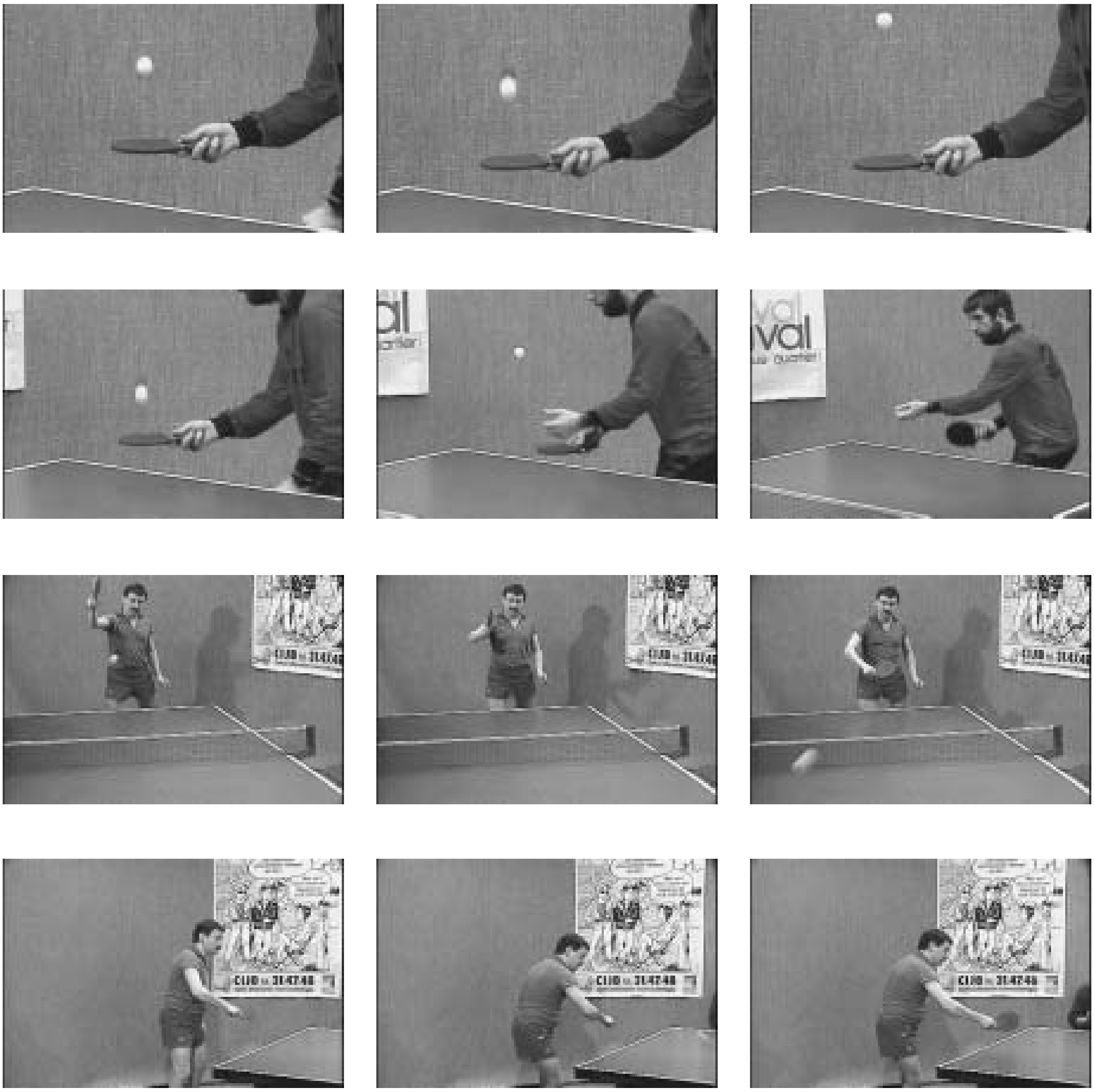

Figure 18: *Table Tennis* sequence: (a) first-row (top): Frames 1, 10 and 20 from left to right (object motion), (b) second-row: Frames 40, 50 and 60 (camera *zoom-out*), (c) third-row: Frames 100, 110 and 120 (object motion), and (d) fourth-row: Frames 200, 210 and 220 (camera *pan-right*)

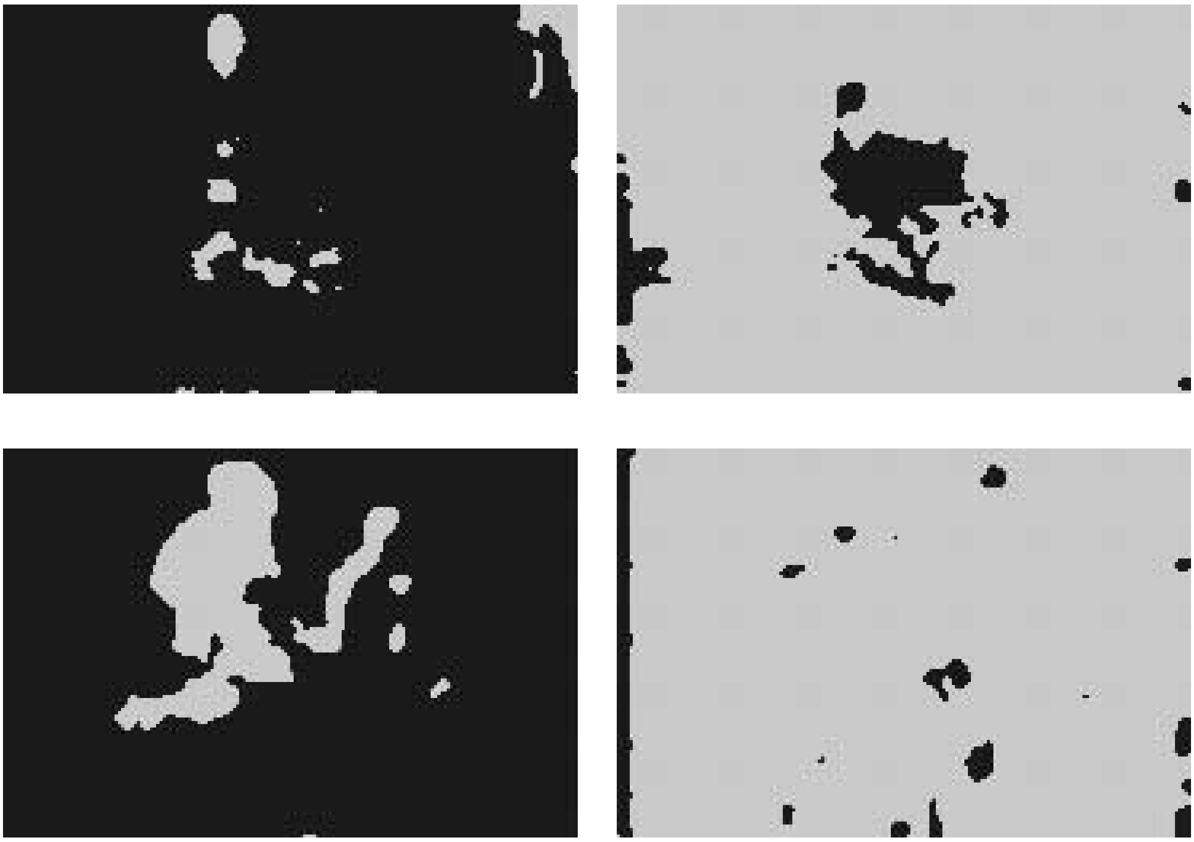

Figure 19: *Table Tennis* sequence: OFS magnitude signatures (thresholded and binarized) measured on Frame 11 (top-left), Frame 41 (top-right), Frame 121 (bottom-left) and Frame 201 (bottom-right)

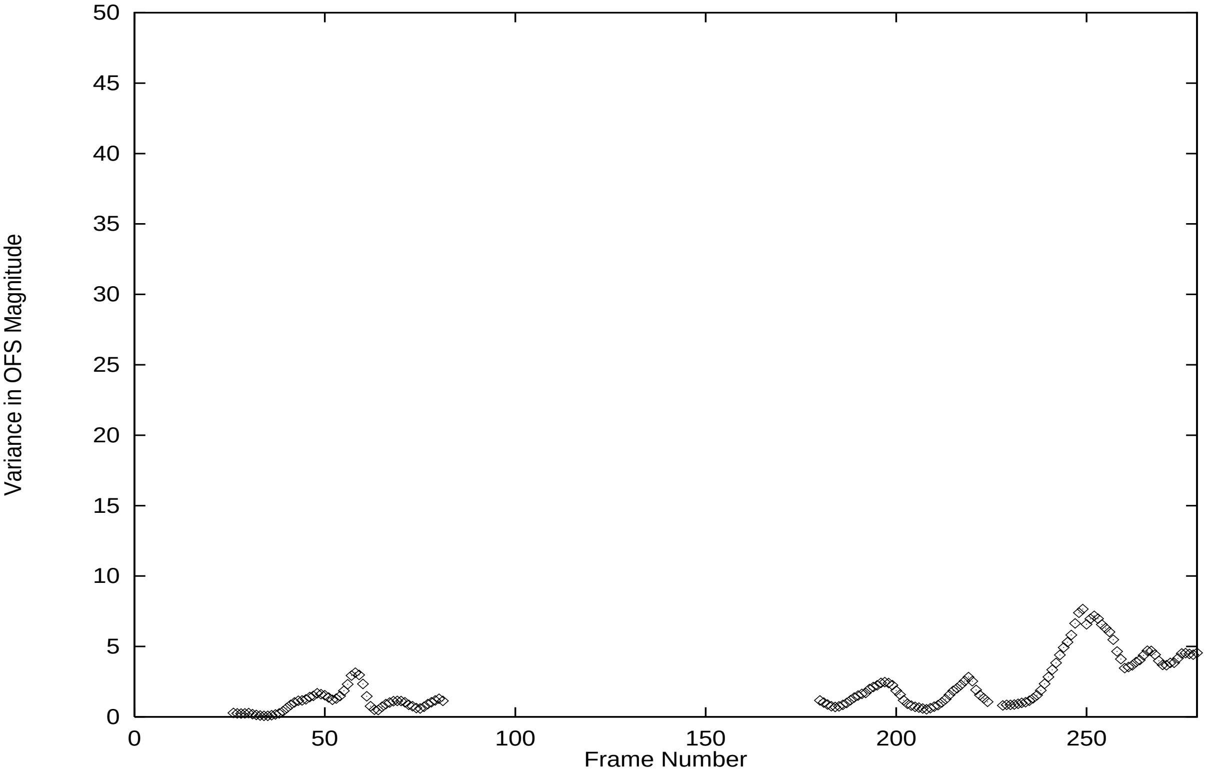

Figure 20: *Table Tennis* sequence: *Singular* and *Non-Singular* variances in the distribution of OFS magnitudes in video frames

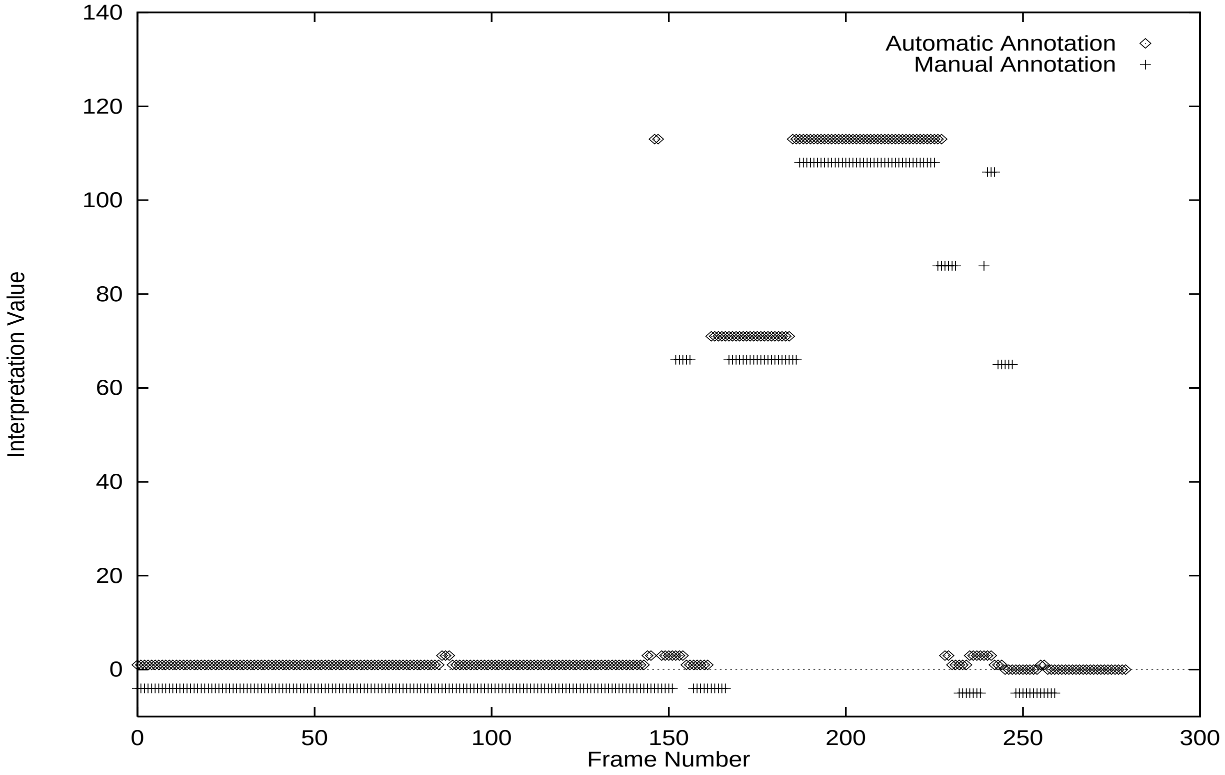

Figure 21: Motion annotation of *Foreman* sequence

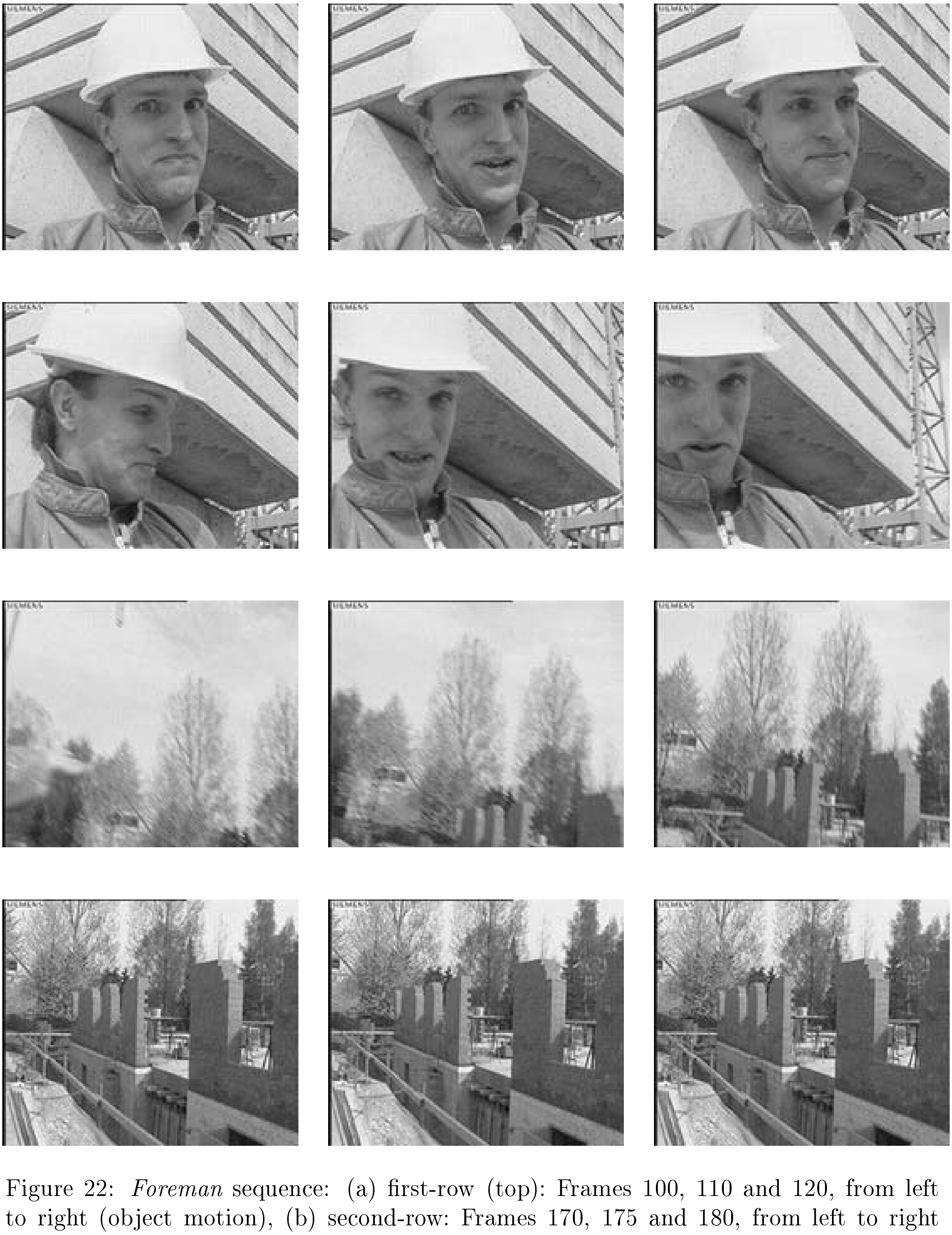

Figure 22: *Foreman* sequence: (a) first-row (top): Frames 100, 110 and 120, from left to right (object motion), (b) second-row: Frames 170, 175 and 180, from left to right (camera *pan-right*), (c) third-row: Frames 205, 210 and 215, from left to right (camera *pan-tilt-right-down*), and (d) fourth-row: Frames 250, 260 and 270, from left to right (no motion)

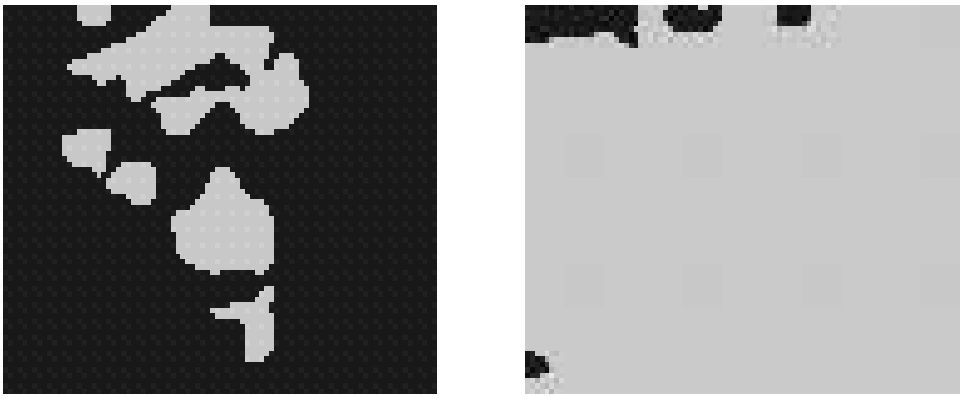

Figure 23: *Foreman* sequence: OFS magnitude signatures (thresholded and binarized) measured on Frame 50 (left) and Frame 200 (right)

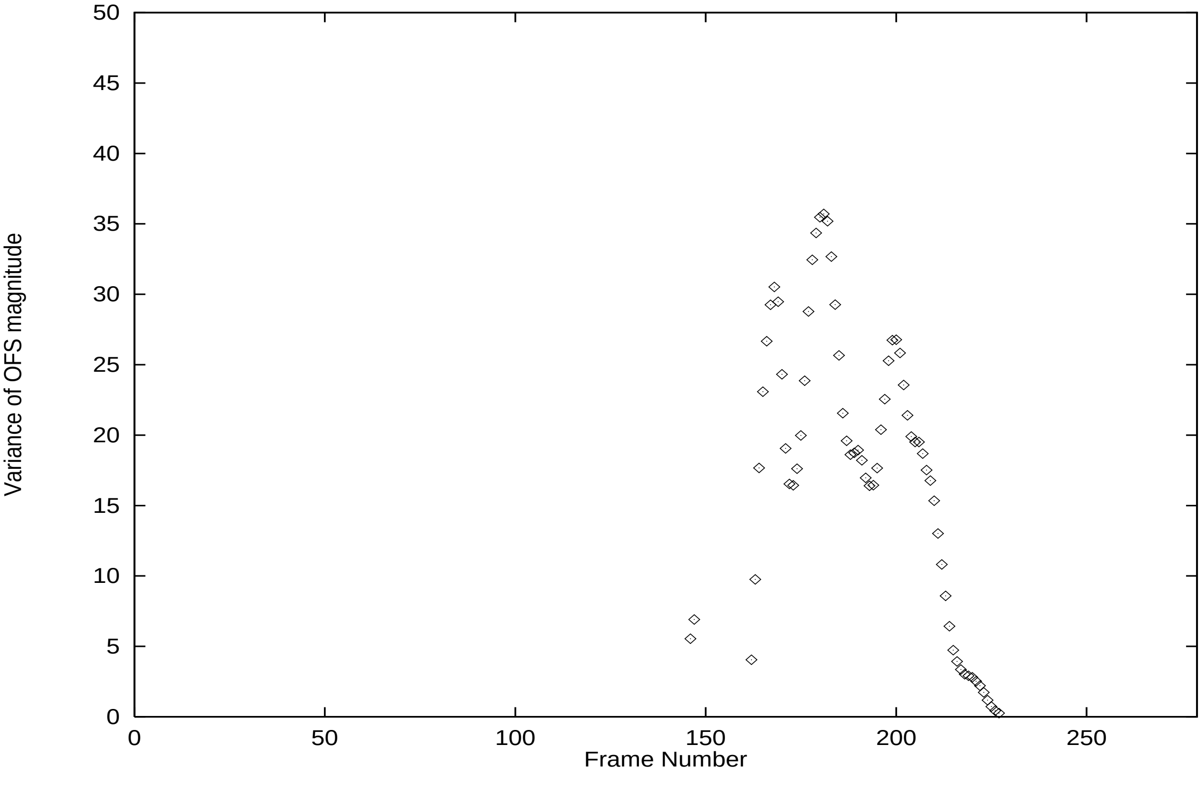

Figure 24: *Foreman* sequence: *Non-Singular* variances in the distribution of OFS magnitudes in video frames

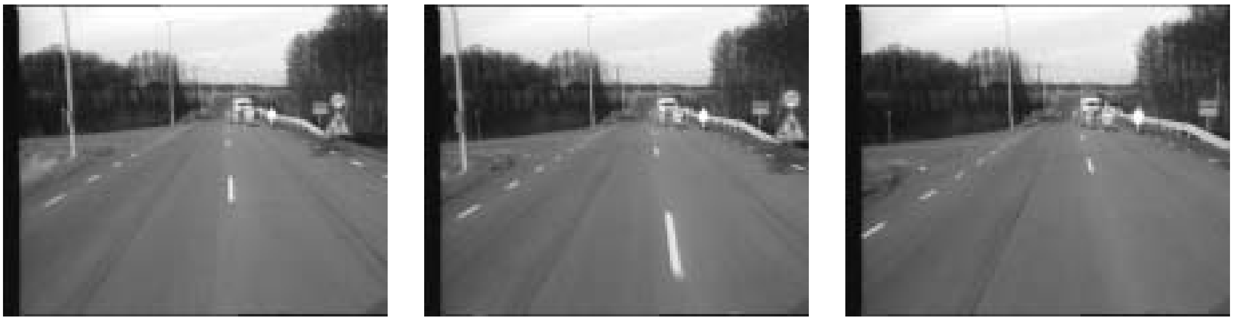

Figure 25: A typical *outdoor* camera *Z-translation* sequence

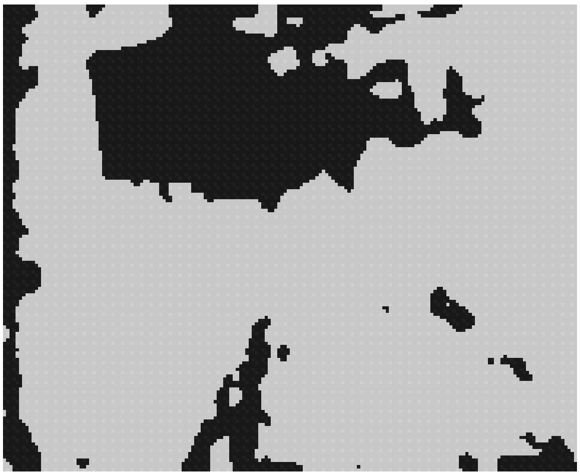

Figure 26: Typical OFS magnitude signature (thresholded and binarized) measured on a frame of the sequence shown in Fig. 25