A Probabilistic Model for Camera Zoom Detection

Rong Jin, Yanjun Qi, Alexander Hauptmann
School Of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

Camera motion detection is essential for automated video analysis. We propose a new probabilistic model for detecting zoom-in/zoom-out operations. The model uses EM to estimate the probability of a zoom versus a non-zoom operation from standard MPEG motion vectors. Traditional methods usually set an empirical threshold after deriving parameters proportional to zoom, pan, rotate and tilt. In contrast, our probabilistic model has a solid probabilistic foundation and a clear, simple probability threshold. Experiments show that this probabilistic model significantly out-performs a baseline parametric method for zoom detection in both precision and recall.

1. Introduction

As digital video becomes more pervasive, effective ways of analyzing video content is increasingly important. Camera operations are one aspect in characterizing a shot to help infer higher-level semantic content [11].

Several approaches have been developed to analyze camera motion of video sequences based on analyzing the optical flow computed between consecutive images [1][4][5]. A few methods directly manipulate MPEG-compressed video to extract camera motion [2][3][6][7]. These approaches use the MPEG motion vectors as an alternative to optical flow.

Intuitively, we know that the motion vectors for a zoom in and zoom out frame will have the "blow out" and "blow in" patterns, respectively. Therefore, in a perfect case, the motion vectors of the macro blocks will point inward or outward to/from the center of the zoom operation with vector length proportional to the distance from the center of the zoom operation.

The difficulty in detecting camera zoom operation comes from noise in the motion vectors due to independent object motion in the frame or the MPEG encoding process, such as quantization errors and other artifacts. Other researchers [5] observed that although the MPEG motion vectors do not represent the true optical flow, they should be sufficient to estimate camera parameters in sequences that do not contain large uniform regions.

In this paper, we propose a novel probabilistic model for detecting zoom in/out camera motion. It is based on the Expectation-Maximization (EM) algorithm for maximum likelihood estimation in the presence of incomplete data. Empirical experiments confirm the superiority of our probabilistic model over a published and high-accuracy parametric method [2][6] of camera motion estimation in terms of precision, recall and F1 score.

1.1. MPEG Motion Vector Field Extraction

MPEG-1 and MPEG-2 streams [12] encode one quantized motion vector per block. Though these motion vectors are not directly equivalent to the true motion vectors of a particular pixel in the frame, there are typically hundreds of motion vectors in one frame, sufficient to estimate camera model parameters. In this work, we use the motion vectors and temporal reference of the P frames extracted from the MPEG-compressed bitstream.

2. A Probabilistic Model for Zoom Detection

Our approach assumes a model where the motion vectors for a frame are produced in part by a perfect zoom and in part by non-zoom noise. In each case, we use Estimation-Maximization to derive the best possible parameters for each of the two models of the motion vectors. We then compare the contribution of the zoom part versus the non-zoom part to the motion vectors for that frame.

The advantage of a probabilistic model for zoom detection is that it naturally handles noisy data. A probabilistic model also produces output probabilities, which are much easier to interpret and utilize than arbitrary threshold values given by typical camera motion detection systems.

2.1. Mathematical Description

In our model, we assume the MPEG motion vector information for the frames is given. The goal of our model is to compute the probability that a given frame is in a zoom-in or zoom-out camera operation. For the i-th frame, let $\vec{v}_i(x, y)$ be the motion vector for the macro block in the x-th column and y-th row and $\mathbf{M}_i$ be our model that explains how all the motion vectors are gener-

ated. To compute the probability for the i-th frame to be from a camera zoom operation, we want to find the optimal model $\mathbf{M}^*_i$ that can best explain the frame's motion vectors and see if that model is consistent with the model of a zoom. In other words, we need to find a model that maximizes the probability of generating all the motion vectors $\{\vec{v}_i(x,y), x=1...m, y=1...n\}$, or

$$\mathbf{M}^*_i = \arg\max_{M_i} P(\{\vec{v}_i(x,y), x=1...m, y=1...n\} \mid \mathbf{M}_i) \quad (1)$$

By assuming each motion vector $\vec{v}_i(x,y)$ is generated independently, we can decompose the conditional probability $P(\{\vec{v}_i(x,y)\} \mid \mathbf{M}_i)$ as

$$P(\{\vec{v}_i(x,y), x=1...m, y=1...n\} \mid \mathbf{M}_i) = \prod_{x=1}^{m}\prod_{y=1}^{n} P(\vec{v}_i(x,y) \mid \mathbf{M}_i) \quad (2)$$

Each motion vector can be generated either due to a zoom-in/zoom-out camera motion or due to something else. Therefore, model $\mathbf{M}_i$ can be decomposed into two parts: $\mathbf{Z}_i$, i.e. the model accounting for zoom camera motion and $\mathbf{N}_i$, i.e. the model for non-zoom reasons. The probability $P(\{\vec{v}_i(x,y)\} \mid \mathbf{M}_i)$ can be rewritten as:

$$P(\vec{v}_i(x,y) \mid \mathbf{M}_i) = P(\mathbf{Z}_i)P(\vec{v}_i(x,y) \mid \mathbf{Z}_i) + P(\mathbf{N}_i)P(\vec{v}_i(x,y) \mid \mathbf{N}_i) \quad (3)$$

Where $P(\mathbf{Z}_i)$ and $P(\mathbf{N}_i)$ stand for the prior probability of using zoom model $\mathbf{Z}_i$ and non-zoom model $\mathbf{N}_i$ to explain the motion vector data, respectively. The probabilities $P(\{\vec{v}_i(x,y)\} \mid \mathbf{Z}_i)$ and $P(\{\vec{v}_i(x,y)\} \mid \mathbf{N}_i)$ are the probabilities of generating the motion vector $\vec{v}_i(x,y)$ using the zoom model $\mathbf{Z}_i$ and the non-zoom model $\mathbf{N}_i$, respectively.

After the general description of the probabilistic model, we need to compute the generation probabilities $P(\{\vec{v}_i(x,y)\} \mid \mathbf{Z}_i)$ and $P(\{\vec{v}_i(x,y)\} \mid \mathbf{N}_i)$. For the non-zoom model $\mathbf{N}_i$, we assume the motion vector $\vec{v}_i(x,y)$ is generated by two Gaussian distributions: one Gaussian distribution generates the amplitudes of motion vectors and the other generates the angles. Let $a_i(x,y)$ and $\theta_i(x,y)$ stand for the amplitude and angle of the motion vector $\vec{v}_i(x,y)$. The probability $P(\vec{v}_i(x,y) \mid \mathbf{N}_i)$ can be written as:

$$P(\vec{v}_i(x,y) \mid \mathbf{N}_i) = P(a_i(x,y) \mid \mathbf{N}_i)P(\theta_i(x,y) \mid \mathbf{N}_i)$$

and

$$P(a_i(x,y) \mid \mathbf{N}_i) = \frac{1}{\sqrt{2\pi\alpha_i^2}} \exp\left(\frac{(a_i(x,y)-\mu_i)^2}{2\alpha_i^2}\right) \quad (4)$$

$$P(\theta_i(x,y) \mid \mathbf{N}_i) = \frac{1}{\sqrt{2\pi\beta_i^2}} \exp\left(\frac{(\theta_i(x,y)-\eta_i)^2}{2\beta_i^2}\right)$$

where μ_i and η_i are the mean amplitude and mean angle of the motion vectors in the i-th frame, α_i and β_i are the standard deviation for amplitudes and angles of motion vectors in the i-th frame.

For the zoom model $\mathbf{Z}_i$, we follow the same idea as for the non-zoom model, i.e. using two Gaussian distributions to describe the generation of motion vectors. The only difference between them is that whereas, in non-zoom model, the mean amplitude and mean angle of motion vectors are position independent, while in the zoom model, these two means are position dependent. As we discussed previously, ideally all the motion vectors should point toward/outward the center of the zoom operation with amplitudes proportional to the distance from the center. Therefore, the mean angle of the motion vector at the position (x, y) would be the angle of the line connecting the position (x, y) and the center (x_c, y_c) and the mean amplitude of motion vector at the position (x, y) should be proportional to the distance between position (x, y) and the center (x_c, y_c). The probability $P(\{\vec{v}_i(x,y)\} \mid \mathbf{Z}_i)$ is

$$P(\vec{v}_i(x,y) \mid \mathbf{Z}_i) = P(a_i(x,y) \mid \mathbf{Z}_i)P(\theta_i(x,y) \mid \mathbf{Z}_i)$$

and

$$P(a_i(x,y) \mid \mathbf{Z}_i) = \frac{1}{\sqrt{2\pi\gamma_i^2}} \exp\left(\frac{(a_i(x,y)-\mu_i(x,y))^2}{2\gamma_i^2}\right) \quad (5)$$

$$P(\theta_i(x,y) \mid \mathbf{N}_i) = \frac{1}{\sqrt{2\pi\delta_i^2}} \exp\left(\frac{(\theta_i(x,y)-\eta_i(x,y))^2}{2\delta_i^2}\right)$$

Where $\mu_i(x,y)$ and $\eta_i(x,y)$ are the mean amplitude and mean angle for the motion vector at position (x, y), γ_i and δ_i are the standard deviation for amplitudes and angles of motion vectors. The values of the mean angle and the mean amplitude for the motion vector at position (x, y) can be written as:

$$\eta_i(x,y) = \tan^{-1}\left(\pm\frac{y-y_c}{x-x_c}\right)$$
$$\mu_i(x,y) = \lambda_i\sqrt{(y-y_c)^2+(x-x_c)^2} \quad (6)$$

Where x_c and y_c are the center of the zoom camera motion. The positive and negative signs within the equation for $\eta_i(x,y)$ will be used for the case of zoom in and zoom out, respectively. In the equation for $\mu_i(x,y)$,

λ_i stands for the proportional constant between the distance and velocity.

The search for optimal model $\mathbf{M}^*_i$ can be divided into two steps. First we fix the center of the zoom motion, i.e. x_c and y_c, and find the parameters μ_i, η_i, α_i, β_i, γ_i, δ_i and λ_i that maximize the probability of generating the motion vector data. Then, varying the center coordinates x_c and y_c to find the best matching center. In the first step, we can employ the Expectation-Maximization algorithm (EM) [8]. The basic idea of EM algorithm is to iteratively update the parameters of the model so that the final model will be guaranteed to be better than the initial model in terms of explaining the motion vectors. For the second step, we can apply the gradient decent algorithm to find the reasonable good matching center coordinates. According to the EM algorithm, the updating equations for parameters μ_i, η_i, α_i, β_i, δ_i, γ_i, λ_i, $P(\{\vec{v}_i(x,y)\}|\mathbf{Z}_i)$ and $P(\{\vec{v}_i(x,y)\}|\mathbf{N}_i)$ are:

$$
\begin{aligned}
\mu_i^{[n+1]} &= \frac{1}{g_i^{[n]}}\sum_{x=1}^{m}\sum_{y=1}^{n} g_i^{[n]}(x,y)a_i(x,y) \\
\alpha_i^{[n+1]} &= \frac{1}{g_i^{[n]}}\sum_{x=1}^{m}\sum_{y=1}^{n} g_i^{[n]}(x,y)(a_i(x,y)-\mu_i^{[n+1]})^2 \\
\eta_i^{[n+1]} &= \frac{1}{g_i^{[n]}}\sum_{x=1}^{m}\sum_{y=1}^{n} g_i^{[n]}(x,y)\theta_i(x,y) \\
\beta_i^{[n+1]} &= \frac{1}{g_i^{[n]}}\sum_{x=1}^{m}\sum_{y=1}^{n} g_i^{[n]}(x,y)(\theta_i(x,y)-\eta_i^{[n+1]})^2 \\
\gamma_i^{[n+1]} &= \frac{1}{z_i^{[n]}}\sum_{x=1}^{m}\sum_{y=1}^{n} z_i^{[n]}(x,y)(a_i(x,y)-\mu_i(x,y))^2 \\
\delta_i^{[n+1]} &= \frac{1}{z_i^{[n]}}\sum_{x=1}^{m}\sum_{y=1}^{n} z_i^{[n]}(x,y)(\theta_i(x,y)-\eta_i(x,y))^2 \\
\lambda_i^{[n+1]} &= \frac{\sum_{x=1}^{m}\sum_{y=1}^{m} z_i^{[n]}(x,y)a_i(x,y)\sqrt{(y-y_c)^2+(x-x_c)^2}}{2\sum_{x=1}^{m}\sum_{y=1}^{n} z_i^{[n]}(x,y)((y-y_c)^2+(x-x_c)^2)} \\
P(\{\vec{v}_i(x,y)\}|\mathbf{Z}_i^{[n+1]}) &= \frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n} g_i^{[n]}(x,y) \\
P(\{\vec{v}_i(x,y)\}|\mathbf{Z}_i^{[n+1]}) &= \frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n} z_i^{[n]}(x,y)
\end{aligned}
\tag{7}
$$

where factors $gi^{[n]}(x,y)$, $g_i^{[n]}$, $z_i^{[n]}(x,y)$, and $z_i^{[n]}$ are computed as:

$$
\begin{aligned}
g_i^{[n]}(x,y) &= \frac{P(\vec{v}(x,y)|\mathbf{N}^{[n]}{}_i)P(\mathbf{N}^{[n]}{}_i)}{P(\vec{v}(x,y)|\mathbf{M}^{[n]}{}_i)} \\
g_i^{[n]} &= \sum_{x=1}^{m}\sum_{y=1}^{n} g_i^{[n]}(x,y) \\
z_i^{[n]}(x,y) &= \frac{P(\vec{v}(x,y)|\mathbf{Z}^{[n]}{}_i)P(\mathbf{Z}^{[n]}{}_i)}{P(\vec{v}(x,y)|\mathbf{M}^{[n]}{}_i)} \\
z_i^{[n]} &= \sum_{x=1}^{m}\sum_{y=1}^{n} z_i^{[n]}(x,y)
\end{aligned}
\tag{8}
$$

The superscript [n] stands for n-th iteration.

Now the probability for the i-th frame to be in a sequence of zoom in/out camera motion can be computed as the averaged probability for motion vectors in the i-th frame to be explained by the zoom model $\mathbf{Z}_i$, or

$$
P_i(zoom) = \frac{1}{mn}\sum_{x=1}^{m}\sum_{y=1}^{n}\frac{P(\vec{v}_i(x,y)|\mathbf{Z}^*{}_i)P(\mathbf{Z}^*{}_i)}{P(\vec{v}_i(x,y)|\mathbf{M}^*{}_i)} \tag{9}
$$

Where $\mathbf{Z}^*{}_i$ is the zoom model within the optimal model $\mathbf{M}^*{}_i$ and $P_i(zoom)$ is the probability for the i-th frame to be in a sequence of camera zoom in/out motion. To decide if a frame is in a sequence of zoom camera motion, we simply set the threshold for zoom probability as 0.5. When the probability $P_i(zoom)$ is larger than 0.5, the frame is retrieved as a frame in a zoom sequence. Otherwise, the frame is decided to be in a non-zoom sequence.

3. Experiments

The effectiveness of a camera zoom detector can be evaluated from two aspects frequently used for information retrieval (IR) [9]:

1) **Precision**: Among the zoom segments detected by the system, how many are truly zoom segments?

2) **Recall**: For all possible zoom segments, how many were found?

A good camera zoom detector should have both high precision and high recall, i.e. a good detection system should retrieve all the zoom segments and nothing else.

If we let *d* be the number of segments detected by the system, *z* be the number of segments manually judged as zoom, and *dz* be the number of segments judged as zoom among the detected segments, the precision and recall for the system are:

$$
\text{Precision} = dz/d \quad \text{Recall} = dz/z \tag{10}
$$

F1 is another common metric used by IR systems that combines precision and recall [9]. It is defined as

$$
\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}
$$

F1 is high only when precision and recall are both high.

3.1. Baseline Model: Threshold-based Parametric Zoom detection

To evaluate our probabilistic method, we compared it to an implementation of the parametric model described in [2][6]. This model uses four parameters to estimate camera motion: zoom, tilt, pan, and rotate. Equation (12) clarifies the relationship between motion vectors (two parameters) and camera operations as:

$$
\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_{zoom} & b_{rotate} \\ -b_{rotate} & a_{zoom} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c_{pan} \\ d_{tilt} \end{pmatrix} \tag{12}
$$

Where $(u\ v)^T$ is a motion vector, $(x\ y)^T$ is the vertical and horizontal position of each frame, a_{zoom}, b_{rotate}, c_{pan}, and d_{tilt}, are scalar coefficients concerned respectively, with zooming, rotation, panning, tilting. The parametric method described in [2] further uses a variation of least-square principle that rejects outliers at each iteration by using a Gaussian distribution to model how well the global motion parameters matches the motion field. When the zoom coefficient is above a global threshold, a frame is classified as part of a 'zoom' camera motion [6].

We carefully implemented the threshold-based parametric system described in [2][6] as our baseline system, since it is typical for traditional, state-of-the-art approaches to camera motion estimation. The global threshold was set based on experiments with training data. Further analysis showed it to be near optimal.

3.2. Experimental Results

To experimentally validate our zoom detection system, we asked a person to mark camera motion operations for 92 different video segments in five different movies from the TREC10 video archive [10]. In addition to zoom in and zoom out camera motions, pan left, pan right, pan up, pan down and rotate camera motions were marked. Out of the 92 segments [about 68.6 minutes], 13 were marked as zoom-out and 13 segments were manually marked as zoom-in segments. We applied both our probabilistic model for detecting camera zooms described earlier and an implementation of the baseline threshold-based parametric model described above to detect zoom in and zoom out segments and then computed precision, recall and F1 scores.

	Zoom In		Zoom Out	
	ProbM	BaseM	ProbM	BaseM.
Precision	0.433	0.267	0.650	0.347
Recall	0.590	0.300	0.667	0.400
F1	0.499	0.282	0.658	0.371

Table 1. Results for zoom detection on 92 segments. "ProbM" is the new probabilistic model and "BaseM" is the traditional model.

According to the results listed in Table 1, the probabilistic model significantly out-performs the traditional model in terms of precision, recall and F1 score in detecting zoom in and zoom out segments.

4. Conclusions

We proposed a novel probabilistic model for detecting zoom in and zoom out camera motion. In our empirical experiment, our probabilistic model significantly outperforms a typical parametric method in terms of precision, recall and F1 score. Three factors contribute to the success of the new probabilistic model:

1) The probabilistic model handles noise better than traditional methods. We introduced a non-zoom model $\mathbf{N}_i$ to account for noise information in the motion vectors.
2) The EM algorithm was used to find an optimal model $\mathbf{M}^*_i$. EM is an algorithm widely used in statistics and is guaranteed to find a local maximum.
3) Universal threshold value. In our model, if the probability $P_i(zoom) > 0.5$ then the frame is determined to be part of a zoom operation. In contrast, the parametric methods have to rely on arbitrary threshold values, which may not be appropriate for a specific video.

Finally we believe our probabilistic model can easily be extended other camera motions, such as pan and tilt, by simply incorporating models of these camera motions.

References

[1] Jinzenji K., Ishibashi S., Kotera H., "Algorithm for automatically producing layered sprites by detecting camera movement", *International Conference on Image Processing* 1997, pp. 767 -770 vol.1

[2] Wang R., Huang T., "Fast camera motion analysis in MPEG domain", *International Conference on Image Processing 1999*, pp. 691 -694 vol.3

[3] Jong-Il Park, Inoue S., Iwadate Y., "Estimating Camera Parameters From Motion Vectors of Digital Video", *IEEE Workshop Multimedia Signal Processing 1998*, pp.105-110

[4] Denzler J., Schless V., Paulus D., Niemann H., "Statistical approach to classification of flow patterns for motion detection", *International Conference on Image Processing 1996*, pp. 517 -520 vol.1

[5] P. Bouthemy, M.Gelgon, and F. Ganansia, "A unified approach to shot change detection and camera motion characterization," *IEEE Trans. Circuits Syst. Video Technology*, no.7, pp. 1030-1044, vol.9, Oct.1999

[6] Ardizzone E., La Cascia M., Avanzato A., Bruna A., "Video Indexing Using MPEG Motion Compensation Vectors", *IEEE International Conference on Multimedia Computing and Systems 1999*, pp. 725 -729 vol.2

[7] Jae-Gon Kim, Hyun Sung Chang, Jinwoong Kim, Hyung-Myung Kim, "Efficient camera motion characterization for MPEG video indexing", *IEEE International Conference on Multimedia and Expo 2000.* pp. 1171 -1174 vol.2

[8] Dempster, A. P. Laird, N. M. Rubin, D. B.. "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, B, pp. 1-38, 1977

[9] Sparck Jones, K. (Ed), "Information Retrieval Experimental", London: Butterworths, 1981.

[10] Voorhees, E. M. and Harman, D. K. *The Tenth Text Retrieval Conference (TREC-10)*, in press, 2001.

[11] Wactlar H.D., Christel M.G., Gong Y., Hauptmann A.G. "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library", *IEEE Computer* 32(2): pp. 66-73.

[12] [MPEG] *Moving Pictures Expert Group*, Standards ISO/IEC 13818-2:2000, and ISO/IEC 11172-2:1993 URL http://mpeg.telecomitalialab.com/standards.htm