Toward an computer-aided sign segmentation

François Lefebvre-Albaret, Frederick Gianni, Patrice Dalle

IRIT (UPS - CNRS UMR 5505)
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 9
{lefebvre,gianni,dalle}@irit.fr

Abstract

The presented article explains an innovating method to process a computer-aided segmentation of sign language sentences. After having tracked the signers hands in a video, the traitment consists in detecting motion attributes such as repetition or symmetries. Those observations are taken into account to process a gesture segmentation. We also discuss about the evaluation of such a segmentation.

1. Introduction

Processing French Sign Language (FSL) videos requires a first segmentation step. Nowadays, this tedious task is processed manually and the result is dramatically influenced by the the human operator. We have focused on the segmentation problem to find unified segmentation criteria and to accelerate the segmentation step. The applications of such a research are far beyond the linguistic task of defining where a sign begins or ends, it could also be applied to automatic sign language video processing or used to produce sign language sentences with signing avatars.

2. Goal of the paper

We first relate some significant studies concerning sign language video processing and present some commonly used algorithms in sign language video processing. Then, we explain the method we have developed to segment a video into signs. This algorithm is based on dynamic programming and on a one-segment definition of a sign. It processes hand motion, and we will soon include other informations as facial expression, elbow position or hand configuration. After having detailed our evaluation method, we will discuss about the accuracy of our segmentation results, and the way we could improve it.

3. Previous studies of sign language video processing

Nowadays, most teams focus on the sign recognition problem. The recognition process sometimes includes a segmentation step (Kim et al., 2001), but the segmentation results are not evaluated. However, those recognition methods are based on several approaches that could also be used for sign segmentation.

The sign recognition methods can be classified into several categories according to the model of sign they refer to. We will distinguish approaches using one-segment or multi-segment sign modelling and other hidden model based algorithms.

In one-segment approach, each gesture is modeled as one single segment. This description refers to Stokoe's sign definition and models a sign as a combination of simultaneous features (hand motion, position, configuration and orientation) (W. C. Stokoe and Croneberg, 1978). The reasons for such a one-segment model of a sign have been exposed in (Channon, 2002). This one-segment model approach has been used in (Derpanis et al., 2004) in order to characterize isolated gestures. Each gesture is qualified by its motion pattern, its hand configuration, and its location. The purpose of their algorithm is to recognise the primitive combination. There are 14 different movements, 3 body locations and 4 hand shapes. Each primitive can be identified with one or more operators processing the whole video sequence of one elementary gesture. According to the operator results, it is possible to determine which combination of primitives has been used to create the gesture. 148 Movements were correctly classified with a success rate of 86 %.
This method has only been applied to gesture classification and was not employed to process real signs ; but this kind of approach could also be useful in sign language processing.

An other approach would be to consider signs as a succession of segments according to the model proposed by Lidell and Johnson (Liddell and Johnson, 1990). This approach has been successfully employed by (Kim et al., 2001) for sign recognition. The algorithm models a sign as a succession of 5 states (resting state, preparation state, stroke/moving state, ending/repetition state and end state) and uses a Markov Model to segment the signs. After this step, a Hidden Markov Model (HMM) leads to sign recognition. The algorithm is able to recognise signs with an accuracy of 95 %. Unfortunatly, there are very fiew information about the evaluation protocol.

In fact, HMM are commonly used for sign recognition because this model is particularly adapted for temporal signal processing. In such a method, each sign is modelled as a succession of states that are automatically determined during the training phase. Processing signs with HMM does not need any *a priori* explicit sign model, but needs a long training phase to find the optimal states. This method is sometime adapted to process the different sign parameters (hand motion, configuration) separately (Vogler, 2003) or to speed up the recognition phase (Wang et al., 2001)[1].

[1]These studies make use of a cyberGlove to capture motion

In the field of sign language video processing, several studies also used HMM based methods. Among them (Bauer and Hienz, 2000) uses HMM to process isolated sign recognition. The signer wears coloured gloves, both hands are then easier to track in the video. The recognition rate is 92 % on a 97 sign corpus.
A very interesting study has been realised by (Bowden et al., 2004). As in the previous studies, hands are tracked to find out in each frame of the video sequence their position and their global shape among 5 configurations. The use of a combination of HMM and Principal Component Analysis enable a recognition rate of about 98% over a 43 sign corpus. Those results are very encouraging but the recognition process only works on isolated signs.

We have presented a few studies related to our present research. Some other methods have been achieved to process isolated and continuous sign recognition. Those approaches are listed in (Ong and Ranganath, 2005).

4. Which approach for a sign language segmentation?

Those ways of sign language processing have been successfully used to perform isolated or continuous sign recognition. Even if a segmentation is made at the same time of the sign recognition, the segmentation accuracy is not evaluated.

The HMM based method requires a long training phase to be able to recognize each sign. Such an approach could not be applied for a continuous natural sign language segmentation (we can notice that all the video used previously for the studies only contained highly constrained sentences with a small set of vocabulary). We have noticed that a lot of signs commonly used in FSL are highly iconic. Moreover, standard signs can be transformed according to the context to express spatial relationships (as it is the case for directional verbs). For those two reasons, learning all the possible signs seems unconceivable with such a method.

The solution would be to detect the sign components independently. This is what has been done by (Vogler, 2003). We have chosen to use this approach to build our segmentation method and to base our algorithm on a one-segment definition of a sign.

5. Algorithm presentation

Our segmentation process is composed of four steps :
Firstly, the hands and head position are tracked in a video.
Secondly , a human operator picks out one frame (that will be called *seed*) of each sign included in the video sequence.
Thirdly, according to the *seeds* and the hands trajectories, the algorithm performs the segmentation.
Fourthly, the sign language expert can check the segmentation result and make the necessary corrections.
Each of those steps depicted in [Figure 1] will be explained in the following sections.

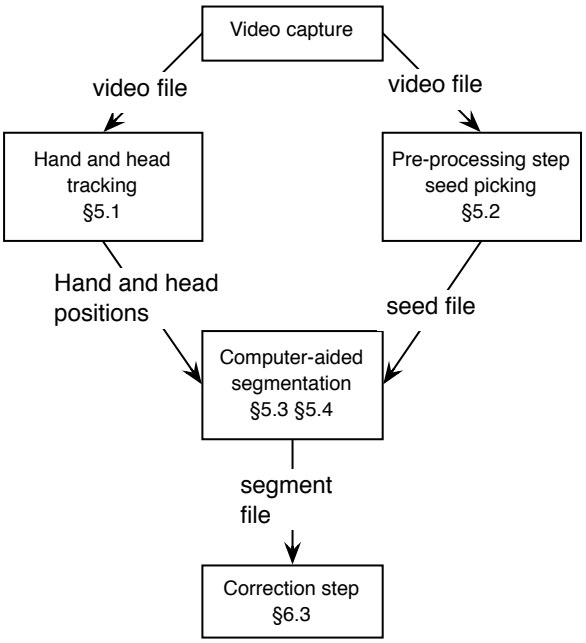

Figure 1: Description a computer-aided segmentation

5.1. Body parts tracking

The first step of the segmentation process consists in tracking the head and the two hands in the video. During a FSL utterance, hand motions can be very fast and brutal direction changes come along. One of the major problem is to design methods that handle those kinds of movements.
In the presented approach, we use the skin color to detect the head and hands, and statistics estimators (via particles filters) for the correspondence. Since particles filter models the uncertainty, it provides a robust framework to track the moving hands of a person telling a story in FSL.

5.1.1. Algorithm description

We used the annealed filtering method presented in (Gall et al., 2006) and applied it in a skin color context, in order to be robust against non-rigid motion and free orientation changes of the observed objects. The observation density is modeled by the skin color distribution of pixels using non-parametric modelling. This model is sometimes referred to as construction of skin probability Map (Brand and Mason, 2000) (Gomez and Morales, 2002).

5.1.2. Results

We have evaluated the tracking method on a FSL video sequence. This sequence is around 3000 frames long (at 25 frames per second) and images have a size of 720 x 576 pixels. The results are presented in the [Figure 2].

5.2. Pre-processing step

By now, a fully automated segmentation, using only motion processing with an unconstrained sign language, gives as a results a 25 % correct segmentation. A short intervention of a human operator in the segmentation process can increase this correct segmentation rate by manually selecting one frame (and only one) of each sign

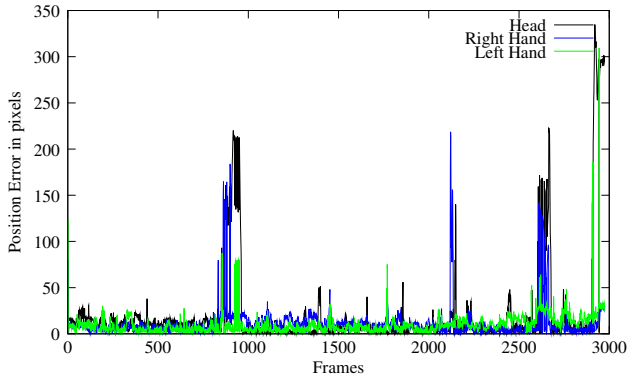

Figure 2: Evaluation the particles filter applied to the FSL video from LsColin.

included in the video sequence. During this step, the video can be displayed with a normal speed or slowed down according to the operator preference. The selected frame can be anywhere in the sign temporal segment.
Each time he recognizes a sign, the operator simply presses a key of his keyboard. The result of this manual pre-processing step is a list of *seed* frames, which is represented as a track at the bottom of the visualization screen [Figure 3]. Naturally, it is possible to make some corrections and to move back if a mistake has been made while pointing the *seeds*.

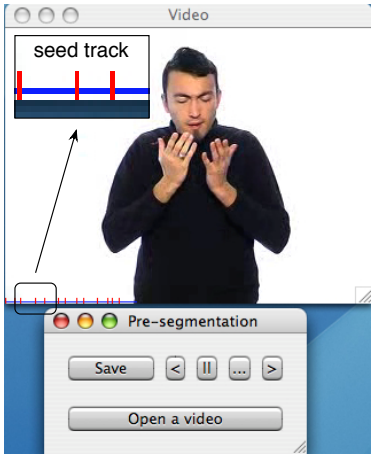

Figure 3: Tool for the pre-segmentation step

5.3. Segment characterisation

Our algorithm is given this list of *seeds* and has to find for each one the beginning and the end of the corresponding sign. The next step consists in processing the hand motion in order to find the temporal segment of each sign that have been marked during the pre-processing step. Some operators characterising a specific feature (motion symmetry, repetition, straight movement) will be applied to each temporal segment. Thanks to those features detections, it will be possible to assign two confidence measures to each temporal segment indicating whether it could be a sign or a transition between two signs. The following part explains how those operators are applied to calculate those confidence measures.

The algorithm processes the 2D hand motion in the video. It means that depth information will not be taken into account. As a consequence, some different motion patterns will be recognised as the same movement. For instance, it will be the case for a horizontal circle and a back and forth horizontal movement.

In the following explanation, a temporal segment between the frame i and the frame j will be noted S_{ij}. The 2D speed of right and left hand on the frame f will respectively be written $\overrightarrow{Vr(f)}$ and $\overrightarrow{Vl(f)}$. The horizontal and vertical components of the right speed will respectively be written $Vr_h(f)$ and $Vr_v(f)$.

Each temporal segment S_{ij} of less than 50 frames is analysed to find out movement features. We use 9 different kinds of operators divided into four categories:

Relational operators :
They detect a specific relationship between the motion of left and right hands during the sign processing:

- Central symmetry : $\overrightarrow{Vr(f)} \approx -\overrightarrow{Vl(f)}$.
- Sagittal symmetry :
  $Vr_h(f) \approx -Vl_h(f)$ and $Vr_v(f) \approx Vl_v(f)$
- Translation : $\overrightarrow{Vr(f)} \approx \overrightarrow{Vl(f)}$
- Static hand (only the case of a static left hand will be illustrated) : during the temporal segment $\frac{||\overrightarrow{Vl(f)}||}{||\overrightarrow{Vr(f)}||} << 1$

Structural operators :

- Double repetition (the global movement can be depicted as a juxtaposition of two identical movements)
- Triple repetition

Motion operators :

- Straight motion (constant speed)
- Straight motion (accelerated and decelerated)

Economy operators :

- A last operator is applied to evaluate the economy of the sign. Assuming the hand to be a punctual mass ($m = 1$). The temporal segment is characterized by the hand speed ($\overrightarrow{V_b}$ and $\overrightarrow{V_e}$), and the hand position (P_b and P_e) at the beginning and at the end of the segment and the duration d of the segment. The minimal energy to realize the transition between the states (V_b,P_b) and the states (V_e,P_e) is written E_m. This energy is calculated thanks to a potential energy difference.
  The real energy is written E_r.

$$E_r = \sum_{f=i}^{j-1}(|V_h(f).(V_h(f) - V_h(f+1))| + |V_v(f).(V_v(f) - V_v(f+1))|)$$

The economy of a movement is computed by the formula E_m/E_r. This operator is only used to find out the transition segments. If the result is near 0, it means that the movement is not economical and probably contains a sign.

Not all of the movements have been taken into account. For instance, we do not detect yet curved trajectories or back and forth motion. Those detectors of movement primitives will be added later if they increase the segmentation accuracy.

The results of those k operators are written o_{ij}^k ($o_{ij}^k \in]0,1]$). An aggregation process, based on weighting rules is applied to assign two confidence measures to each temporal interval. The first one, C_{ij}^s, indicates weather the segment could be a sign and the second one, C_{ij}^t, indicates weather it could be a transition. The aggregation process will be improved and studied in detail in the further version of our algorithm.

5.4. Dynamic programming

Afer having given those two confidence measures to each temporal segment, we want to find the best video segmentation. All sign sentences are modelled by a succession of signs and transitions. The structure of a sign sentence is then represented as the following finite state machine [Figure 4].

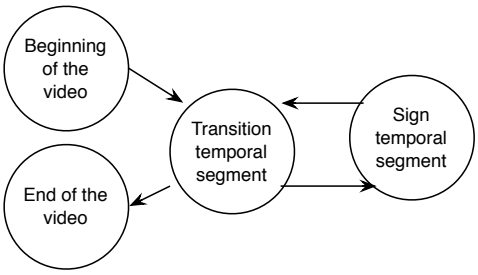

Figure 4: Sentence model

We can draw a parallel with a conventional two states Markov Chain. Each state does not represent a time point but is actually a temporal segment (a succession of frames). We have adapted Viterbi algorithm, which is currently used to solve this kind of problem to take the temporal segment length into account.

The function that should be optimized to find the segmentation is then :

$$argmax\left(\sum_n ln(C_{ij}(n)).l(n)\right)$$

were $C^{ij}(n)$ and $l(n)$ represent the confidence measure and the length of the n^{th} chosen temporal segment.
Other constraints are also added to use the *seeds* :

- All signs must contain exactly one *seed*
- A transition have not to contain any *seed*

The result of the segmentation is a succession of segments, which can be visualized with the annotation software Ancolin (Braffort et al.,). This software allows us to compare the computer-aided segmentation with a manual segmentation [Figure 5].

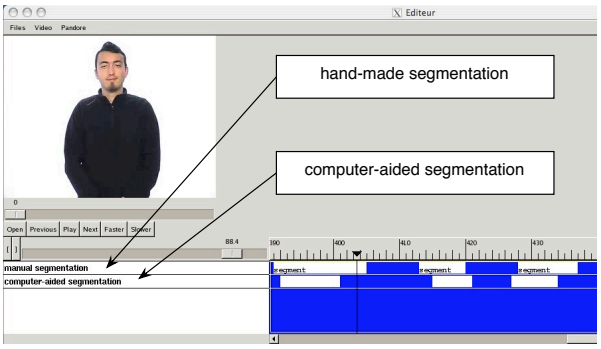

Figure 5: Comparison of manual and computer-aided segmentation

6. Result evaluation

6.1. Evaluation criteria

How can we evaluate a segmentation? The answer mainly depends of the usage of the obtained temporal segments. Our mid-term goal is to animate a signing avatar. As a consequence, the segmented signs are intended to be used in other sentences. If they do not contain the whole definition of the sign, a truncated sign will always be displayed. At the opposite, if a sign contains a part of the transition from the previous or to the following sign, this parasit motion or configuration will degrade the quality of the sign language synthesis.

In general, partially segmented signs present one of the following features :

- Some contacts or a speed reversal points of the original sign are not included in the segment.
- The number of repetitions or back and forth cycles is not correct .
- The segment contains a configuration of the preceding or following sign.
- The segment contains a movement of transition from the preceding or to the following sign.

6.2. Result analysis

We have used two sequences to test our algorithm. It is very important to notice that those two sequences have been realised by native signers and that the signs are processed at a normal speed. At the contrary of a lot of other studies, the vocabulary used in the video is totally unconstrained.
The first one is a 2 minutes video sequence where the signer explains the 11 September 2001 events from LS Colin corpus (Cuxac, 2002). This sequence is very interesting because a lot of signs are highly iconic. The story is composed of 203 signs. Our success rate is 40% in this video.
The second video[2] has been provided by the Websourd society and is a translation of a piece of news in FSL. This video of 30 seconds contains 43 signs. Among them :

- 22 are correctly segmented (54%)
- 16 are partially segmented (39%)
- 3 are not detected at all (7%)

It clearly appears that the segmentation accuracy depends of the type of the processed video. The following hypothesis could account for this accuracy difference between the two computer-aided segmentations.

Video length : For small videos, it is possible to watch the video once. The seeds picking is then easier to realise because one can remember the sign included in the video. It was not possible for the 11 September video because the video was to long. As a consequence, a lot of seeds have been misplaced during the pre-processing step.

Prepared sentences : The 11 September story was totally spontaneous and had not been prepared. So, there was a lot of hesitation that we have considered as if it were signs. Unfortunately, a lot of them were very short and did not fit our automatic segmentation criteria.

Iconic signs : A lot of iconic signs have been used in the 11 September video. For some of them, it was very hard to distinguish elementary sub-signs, which could be automatically segmented. For this reason, we had some granularity problems. Some signs, that were considered as one segment during the hand-made reference segmentation, were assigned two seeds during the pre-processing step (seed-picking) because they seemed to be compound of two elementary signs.

Coming back to our goal of signing avatar animation, the errors made for hesitation and highly context dependant signs are not so penalising because those signs would be hard to reuse to build other sentences. However, such signs have to be taken into account in a computer-aided segmentation tool to be able to process real sign language.

The partially segmented signs of the second video (news translation) have been classified into several categories :

- 1 segment had too little frame. The sign was not recognizable.
- 6 segmented signs were truncated (one contact point, repetition or back and forth motion have been deleted).
- 5 segments contained a configuration of the previous or following sign.
- 4 segments contained a movement of the previous or following sign (or a movement of the transition to those signs).

Those results could be improved by taking other parameters into account in the segmentation process. According to our observations, the hand configuration exploitation would be a good way of improving the segmentation accuracy.

6.3. Correction step

Even if the above results are very encouraging, automatic segmentation must be checked and corrected by a human operator. During this step, all the segment must be visualised and corrected if needed.

It is very important to avoid deletion errors, because the presence of ignored signs would oblige the sign language expert to watch all the video to find out those signs.

It would also be interesting to use the confidence measure C_{ij}^{s} during the checking phase. By using this value, the sign language expert would be able to focus on the signs, which have the smallest confidence value (and then the highest probability to be bad segmented).

The goal of computer-aided segmentation is to accelerate the manual segmentation phase. It means that the pre-processing step added to the checking step must spend less time than a fully manual segmentation. We have measured those 3 informations for the second video (of 30 second) :

- Seed picking : $T_s = 3$ minutes
- Checking/correction : $T_c = 7$ minutes
- Manual segmentation : $T_m = 13$ minutes

$(Ts + Tc)/Tm = 0.77$. In our case 23 % time is spared using this semi-automated segmentation process. And we could improve this value by increasing the segmentation accuracy of our algorithm.

7. Conclusion

Regarding to the few parameters that have been taken into account to process the segmentation our results are very encouraging and validate our segmentation approach.
However, it is important to improve the segmentation accuracy to make our method be usable in corpora processing. Such an improvement could only be made in taking into account other sign parameters:

[2] This video can be downloaded at the following address http://websourd.nnx.com/ mediav0/IMG/flv/1D001-97.flv

- A lot of sign ends can be characterized by a configuration change. Using configuration could probably allow us to perform a better segmentation.
- The simultaneous tracking of hand, elbow and shoulder positions could lead to a reconstruction of the whole arm posture as demonstrated in (Lenseigne et al., 2004) and allow us to process 3D motion. The hand motion could be depicted more accurately.
- The analysis of head orientation and facial expression could also decrease the number of wrong segmentation (Parashar, 2003).

8. Acknoledgments

This study has been financed by the Websourd society and the Midi-Pyrenees region. It has been done in the frame of the SESCA project (System for Sign Language Pedagogy and Communication with Avatars) of the group PRESTO (Tolosan Sign Reserch Group). Thanks to J. Dalle who has been our LSF expert and performed all the manual segmentation that were refered to in this article.

9. References

Britta Bauer and Hermann Hienz. 2000. Relevant features for video-based continuous sign language recognition. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 440, Washington, DC, USA. IEEE Computer Society.

Richard Bowden, David Windridge, Timor Kadir, Andrew Zisserman, and Michael Brady. 2004. A linguistic feature vector for the visual interpretation of sign language. In Tomás Pajdla and Jiri Matas, editors, *ECCV (1)*, volume 3021 of *Lecture Notes in Computer Science*, pages 390–401. Springer.

A. Braffort, A. Choisier, C. Collet, P. Dalle, F. Gianni, B. Lenseigne, and J. Segouat. Toward an annotation software for video of Sign Language, including image processing tools & signing space modelling .

Jason Brand and John S. Mason. 2000. A comparative assessment of three approaches to pixel-level human skin-detection. In *15th International Conference on Pattern Recognition*, volume 1, pages 1056–1059.

R.E. Channon. 2002. *Signs are single segments: phonological representations and temporal sequencing in ASL and other sign languages*. Ph.D. thesis, University of Maryland, USA.

Cuxac. 2002. "langage et cognition" , rapport de fin de recherche. http://www.irit.fr/LS-COLIN.

Konstantinos G. Derpanis, Richard P. Wildes, and John K. Tsotsos. 2004. Hand gesture recognition within a linguistics-based framework. In *Computer Vision - ECCV 20048th European Conference on Computer Vision*. Springer.

J. Deutscher, A. Blake, and Reid I. 2000. Articulated body motion capture by annealed particle filtering. *Computer Vision and Pattern Recognition*.

J. Gall, J. Potthoff, C. Schnoerr, B. Rosenhahn, and H.P. Seidel. 2006. Interacting and annealing particle filters: Mathematics and a recipe for applications. Technical Report MPI-I-2006-4-009, Max-Planck Institute for Computer Science.

G. Gomez and E. Morales. 2002. Automatic feature reconstruction and a simple rule induction algorithm for skin detection. In *ICML Workshop On Machine Learning in Computer Vision*, pages 31–38.

J.-B. Kim, K.-H. Park, W.-C. Bang, J.-S. Kim, and Z. Bien. 2001. Continuous korean sign language recognition using automata based gesture segmentation and hidden markov model. In *ICCAS2001*, Jeju, Korea. Jeju National University.

Boris Lenseigne, Frédérick Gianni, and Patrice Dalle. 2004. Estimation mono-vue de la posture du bras en utilisant un modèle biomécanique, méthode et évaluation. In *14ème Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle - RFIA 2004, Toulouse, 28/01/04-30/01/04*, volume 2, pages 957–966. AFRIF-AFIA, janvier.

S. K. Liddell and R. E. Johnson. 1990. American sign language: the phonological base. *Sign Language Studies*, 64.

Sylvie C.W. Ong and Surendra Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891.

Ayush S. Parashar. 2003. *Representation and interpretation of manual and non-manual information for automated Amercan Sign Language recognition*. Ph.D. thesis, University of South Florida, USA.

C. P. Vogler. 2003. *American sign language recognition: reducing the complexity of the task with phoneme- based modeling and parallel hidden markov models*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA. Supervisor-Dimitris N. Metaxas.

D. Casterline W. C. Stokoe and C. Croneberg. 1978. *A dictionary of American Sign Language on Linguistic principles*. Linstok Press.

C. Wang, W. Gao, and Z. Xuan. 2001. A real-time large vocabulary continuous recognition system for chinese sign language. In *PCM '01: Proceedings of the Second IEEE Pacific Rim Conference on Multimedia*, pages 150–157, London, UK. Springer-Verlag.