EVALUATION OF VIDEO SEGMENTATION METHODS FOR SURVEILLANCE APPLICATIONS

Kevin McKoen[*], *Raquel Navarro-Prieto*[*], *Benoit Duc*[†], *Emrullah Durucan*[‡], *Francesco Ziliani*[‡], *Touradj Ebrahimi*[‡]

[*]Motorola Labs UK, Basingstoke, England.

[†]Motorola inc., Semiconductor Products Sector, Geneva, Switzerland.

[‡]Swiss Federal Institute of Technology, Lausanne, Switzerland.

e-mail: Kevin.McKoen@motorola.com

ABSTRACT

Current research on performance evaluation of video segmentation methods is primarily focussed on the development of objective figures of merit. There is no standardised methodology for subjective evaluations of segmentation performance and these are currently perceived as too onerous.

Using an experimental design and data analysis method derived from current practice in experimental psychology, we have explored a performance evaluation procedure that is largely based on subjective assessments.

We report on statistically significant differences in perceived performance between three multi-object video segmentation and tracking methods developed for surveillance applications. The assessments were performed by a group each of experts and novices on a wide range of video content.

1 INTRODUCTION

The "Segmentation and Tracking for Surveillance Applications" project has been a collaborative effort between the Signal Processing Laboratory of the EPFL, Motorola and Siemens under the Swiss government "CTI" programme, number 3502.1.

This paper reports on the results obtained from the evaluation of three candidate segmentation algorithms that were developed within the framework of this project. The evaluation has combined the efforts of segmentation algorithm designers, video processing engineers and cognitive and experimental psychologists.

A trial algorithm evaluation, carried out in 1999, suggested that:

a) more effective segmentation representations were needed;

b) there were important psycho-visual aspects to be considered in performing subjective segmentation assessments;

c) because segmentation objectives are application dependent, a clear context description is needed to facilitate the subjective assessments.

Automatic video segmentation is widely regarded as an important enabling technology for multimedia services based on standards such as ISO MPEG-4 and MPEG-7 [2, 4, 10]. To date, research on the performance evaluation of video segmentation methods has primarily focussed on the development of objective figures of merit [2, 3]. This is also the main video segmentation evaluation methodology currently proposed in the COST 211quat "Call for Comparisons" [1]. In contrast, there is, as yet, no standardised methodology for subjective evaluations of segmentation performance.

2 SEGMENTATION ALGORITHMS

The main task of the segmentation algorithms developed in this project was to identify moving and/or novel objects (especially people) and track them reliably. Unlike foreground/background segmentation methods that produce a binary object mask, the segmentation methods in this project were required to segment and track multiple objects individually. A number of novel segmentation methods have been investigated in this framework.

The current three candidate algorithms for the evaluation are all based on an initial statistical change detection step [5]. Algorithm 1 is a low complexity approach that uses a connected-components analysis of the change detection map together with extensive region-based tracking [6].

Algorithms 2 and 3 [7] use an extra incremental change detection step to reduce the effect of changes in ambient illumination. They also include a more sophisticated spatial segmentation step based on efficient multi-feature clustering. Region-based motion estimation is then used to track the extracted regions. Algorithm 3 performs a complete spatial segmentation on the whole image before keeping only those regions that have changed whereas Algorithm 2 only segments the change map.

3 EVALUATION VIDEO SEQUENCES

A varied set of six test video sequences, each lasting approximately 15 seconds, were chosen to evaluate the can-

	Algorithm Evaluation Criteria	Evaluation Method
1	Segments moving 'semantic' objects from the background	Subjective assessment
2	Tracks individual regions throughout the video sequence	Subjective assessment
3	Provides accurate region, or preferably object, boundaries	Subjective assessment
4	Distinguishes between moving objects and image perturbations (e.g. camera noise, rain)	Subjective assessment
5	Segments objects into associated sub-regions	Subjective assessment
6	Eliminates or correctly identifies shadows	Subjective assessment
7	Low computational complexity	Run-time data
8	Has few configuration parameters	Run-time data
9	The segmentation is illumination invariant	Post-assessment analysis
10	The segmentation performs well for outdoor sequences	Post-assessment analysis

Table 1: Evaluation Criteria and Performance Metrics

didate algorithms. These ranged from the relatively simple "Hall Monitor" indoor test sequence used by MPEG-4, to outdoor sequences containing very many persons walking and occluding each other in the field of view. There were also two sequences containing large changes in illumination with and without real moving objects being present. Because none of the segmentation algorithms included global motion compensation, all of the test sequences contained a static background.

4 EXPERIMENTAL DESIGN OF THE SUBJECTIVE ASSESSMENTS

4.1 Experimental Design

A Two variable Mixed design (Algorithm x Sequence Type) with repeated measures was chosen for this experiment. The Algorithm variable was manipulated within subjects (i.e. each of the subjects viewed all three Algorithms). The Sequence Type was manipulated between subjects with 6 levels (sequence a, b, c, d, e, and f) and the sequences were classified into three key sequence dimensions. Sequences a, b, c, d were classified by the crossing of the 'Simple Object/Many Objects' and 'Indoors/Outdoors' dimensions. Sequences e and f were classified into the 'Illumination Changes' dimension. The order of appearance of the algorithms was balanced across the subjects.

4.2 Procedure

Each subject saw three sequences which represented three conditions of the Algorithm by Sequence Type design. For each sequence, the subjects saw the original unsegmented sequence as many times as they wanted to, for a maximum time of three minutes. Then the subjects were presented with the entire segmented sequence once. After that, they saw the first third, the second third and the last third of the sequence consecutively. During each part, the subjects completed an assessment rating form. The maximum time for viewing and assessing each sequence part was 5 minutes.

4.3 Measures

The subjects were asked to complete assessment forms with questions that addressed the subjective evaluation criteria specified in Table 1. The ratings were given on a 5 points scale. After pilot assessments, the number of ratings was reduced from 6 to 4. Criteria 1 and 4 were combined and criteria 5 and 6 were also combined.

4.4 Subjects

24 subjects participated in this evaluation. Half of them were expert video users. These were subjects who either used video monitoring systems in their work or had technical video encoding or editing experience. The remaining subjects had no experience of this kind. Since colours were used to represent the segmented areas, all subjects were filtered by a Colour Blind test.

4.5 Viewing Conditions

The assessments were performed in quiet surroundings using high quality computer monitors and viewing distances as recommended in ITU-T Recommendation P.910 [8].

4.6 Validation of Segmentation Representation

The segmented regions generated by the segmentation algorithms were represented by the use of colours. A preliminary study was conducted to evaluate the best representation and to anticipate any problems with this method of visualising the segmentation results. The study confirmed that subjects found the colour coded segmented regions to be differentiable.

4.7 Statistical Analysis used

An ANOVA analysis of variance method [9] was used to extract statistically significant differences between mean ratings. This approach partitions the total variance into the component that is due to the true random error (i.e. within-group variability) and the components that are due to differences between means (i.e. between-group variability). These latter variance components are tested for statistical significance. If these are significant, the null hypothesis (there are no differences between the means) is rejected and the alternative hypothesis (the means are different from each other) is accepted. The significance threshold for this test was set at 5%, as is customary in these experiments.

5 SUBJECTIVE ASSESSMENT RESULTS

The principal goal of the subjective evaluation was to identify the weaknesses and strengths of the project's current segmentation algorithms.

Overall, our data did not provide statistically significant evidence that any one of the candidate algorithms was rated the highest over all the criteria and all the sequences. This was probably because the algorithms were tested over a significant range of sequences and for several quite different evaluation criteria.

At a more detailed level however, individual algorithms were rated higher, on average, for certain criteria. **Algorithm 3**, with the more sophisticated spatial segmentation and refinement of extracted regions techniques, did produce noticeably improved ratings for *the extraction of significant objects* and for *the partitioning of objects into sub-regions.* The extra incremental change detection (i.e. change relative to the previous frame) may also have helped improve ratings on a sequence containing strong changes in ambient illumination but no moving objects. The extensive region-based tracking methods implemented in **Algorithm 1** were reflected in higher ratings for *the consistent tracking of extracted objects.*

On average, all the algorithms were rated highest for *the extraction of significant objects* (rather than the subsequent accuracy of those objects or their consistent tracking). It is also important to note that, on average, the algorithms were rated significantly higher on the sequences with simple objects than those containing complex multiple-object movements.

Regarding criterion 9, "The segmentation is illumination invariant", all algorithms were rated lowest on an indoor sequence with strong illumination changes. For criterion 10, "The segmentation performs well for outdoor sequences", there was no statistically significant difference between the rated performance on outdoor and on indoor sequences.

6 OBJECTIVE PERFORMANCE DATA

6.1 Algorithm Parameters

A distinction was made between algorithm parameters that could be fixed at design time and those that were varied to maintain segmentation performance for different test video sequences. The low complexity Algorithm 1 was able to process the six sequences without requiring any run-time parameters.

Algorithms 2 and 3 identified the image noise estimator in the change detection as a potentially important parameter to vary. This may have helped to improve the perceived performance of Algorithms 2 and 3 on this sequence compared with Algorithm 1.

Algorithm 1	Algorithm 2	Algorithm 3
0.52 sec	29.9 sec	35.8 sec

Table 2: Execution times of the candidate segmentation algorithms

6.2 Computational Complexity

The computational complexity of the candidate algorithms was estimated by comparing their execution speeds over several representative frames of the "Hall Monitor" sequence in CIF format. Execution speeds were measured with the Unix 'time' utility on a Sun Sparc Ultra 5 computer and are presented in Table 2. Although Table 2 indicates that the execution of Algorithm 1 is currently 60-70 times faster than Algorithms 2 and 3, some of this difference may be due to the software implementation of the algorithms. More specifically, Algorithm 1 contained some optimised routines and Algorithms 2 and 3 contained routines written in C++. Nevertheless, it seems likely that both Algorithms 2 and 3 are inherently at least 10 times more complex than Algorithm 1.

7 CRITIQUE OF THE EVALUATION METHOD

An important goal of this work was to research and validate a methodology for assessing automatically generated segmentation maps.

The evaluation has emphasised the need to test segmentation algorithms with a significant range of video sequence content, as we were careful to do, in order to avoid sequence dependent assessments.

We also discovered a difference in the ratings given by our novice and expert subjects. Experts considered that the algorithms performed poorly in *the partitioning of objects into sub-regions.* Novices considered that the poorest performance of the algorithms was in *the consistent tracking of extracted objects.*

The experimental design approach and ANOVA data analysis technique allowed us to focus on statistically significant differences in rated performance.

Some of the subjects were not able to provide ratings of object segmentation performance on a sequence that contained no moving object. In future, this type of ambiguity has to be anticipated and addressed in the design of assessment materials and procedures.

Another key experimental aspect to consider is the selection of the variables to be balanced or blocked. This evaluation balanced the algorithms and their order of presentation and blocked the sequence type (Indoors, Outdoors, Illumination Change).

If it were possible to simplify the ratings and reduce their duration (e.g. by reducing the number of assessment criteria), each subject could assess all algorithms over all the sequences.

Explicit reference segmentation masks were not used in this evaluation. Instead, our segmentation representation allowed the original video sequence to be viewed in monochrome beneath coloured segmented regions. Viewers were therefore able to see the moving objects and form their own, implicit, reference segmentations. This had the advantage of allowing users to determine individually which moving objects were significant or not.

8 COMPARISON WITH COST 211Q SEGMENTATION ASSESSMENT

Video coding researchers are familiar with the need to perform conclusive video quality assessments using subjective evaluations, even if objective measures such as PSNR are widely used during algorithm development. To what extent is this also true for the assessment of video segmentation performance?

The de-facto standard for objective measurement of segmentation mask quality has been developed by Villegas and Marichal [2] within the framework of the COST 211q "Call for Comparisons". This method computes a number of difference measures (currently spatial accuracy, temporal coherence and tracking consistency) between the segmentation mask under assessment and a reference mask. The advantages of this approach are that:

1. it provides detailed and time-dependent data on video segmentation performance;
2. it can be computed in a time-efficient and reliable manner.

On the other hand, our experience with a subjective evaluation method suggests that the subjective evaluation approach provides some different benefits.

1. It has often been remarked [2, 4, 10] that segmentation performance is application dependent. In subjective evaluations, this context may be stated explicitly and experts from this application domain recruited to perform assessments.
2. With a suitable multi-object representation, multi-object segmentation assessments are relatively straightforward for subjects to perform. Viewers can be left to judge how many distinct objects there *really* are in a complex scene and compare this with the objects proposed by the segmentation.

9 CONCLUSIONS

We have reported on a subjective evaluation method for multi-object segmentation performance assessment that has successfully highlighted subtle differences in perceived performance between three candidate segmentation algorithms.

Taking into account the different benefits of segmentation assessment by objective measures [2] and by subjective ratings, we would like to suggest that further comparison between the two approaches is required.

References

[1] COST 211quat "Call for Comparisons" at http://www.teltec.dcu.ie/cost211/

[2] P. Villegas, X. Marichal and A. Salcedo, "Objective evaluation of segmentation masks in video sequences", WIAMIS'99 workshop, Berlin, May 1999.

[3] M. Wollborn and R. Mech, "Procedure for Objective Evaluation of VOP Generation Algorithms", Doc. ISO/IEC JTC1/SC29/WG11 MPEG97/2704, Fribourg, October 1997.

[4] R. Castagno, T. Ebrahimi and M. Kunt, "Video segmentation based on Multiple Features for interactive multimedia applications", IEEE Transactions on Circuits and Systems for Video Technology, special issue on "Image and video processing for emerging interactive multimedia services", vol. 8, no. 5, pp. 562-571, September 1998.

[5] T. Aach, A. Kaup and R. Mester, "Statistical model-based change detection in moving video", Signal Processing, vol. 31, pp. 165-180, 1993.

[6] B. Duc, S. Brunetton, S. Soudagar and K. McKoen, "Motion-based segmentation for wireless surveillance applications", COST 254 Workshop on Intelligent Communication Technologies and Applications with Emphasis on Mobile Communications, Neuchatel, Switzerland, May 1999.

[7] F. Ziliani and A. Cavallaro, "Image analysis for video surveillance based on spatial regularization of a statistical model-based change detection", ICIAP'99, pp.1108-1111, Venezia, Italy, September 27-29, 1999.

[8] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications", August 1996.

[9] H.R. Lindman, "Analysis of variance in complex experimental designs", San Francisco: W. H. Freeman & Co.,1974.

[10] P. Salembier and F. Marques, "Region-Based Representations of Image and Video: Segmentation and Tools for Multimedia Services", IEEE transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1147-1169, December 1999.