A User-Centered Approach to Rushes Summarisation Via Highlight-Detected Keyframes

Daragh Byrne, Peter Kehoe, Hyowon Lee, Ciarán Ó Conaire, Alan F. Smeaton,
Noel E. O'Connor and Gareth J.F. Jones
Centre for Digital Video Processing (CDVP) & Adaptive Information Cluster
Dublin City University,
Glasnevin, Dublin 9, Ireland
+ 353 1 700 5262
alan.smeaton@computing.dcu.ie

ABSTRACT
We present our keyframe-based summary approach for BBC Rushes video as part of the TRECVid Summarisation benchmark evaluation carried out in 2007. We outline our approach to summarisation that uses video processing for feature extraction and is informed by human factors considerations for summary presentation. Based on the performance of our generated summaries as reported by NIST, we subsequently undertook detailed failure analysis of our approach. The findings of this investigation as well as recommendations for alterations to our keyframe-based summary generation method, and the evaluation methodology for Rushes summaries in general, are detailed within this paper.

Categories and Subject Descriptors
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems - *Video (e.g., tape, disk, DVI*); I.2.10 [Vision and Scene Understanding]: Video Analysis

General Terms
Design, Experimentation, Human Factors.

Keywords
Video Summarisation, BBC Rushes, TrecVid

1. INTRODUCTION
In 2007, the National Institute of Standards and Technology (NIST) coordinated an evaluation of a variety of approaches to automatic summarisation of video footage. This took place as part of a larger video benchmarking activity known as TRECVid. The overall video summarisation task, the data used, evaluation metrics, etc., are described elsewhere [9]. In this paper we report our approach and results of completing the summarisation task.

As part of our participation in the TRECVid 2007 video summarisation evaluation we explored the construction of a keyframe-based summary of BBC rushes content. Our summaries were heavily informed by human factors considerations but also made use of our existing work in digital video processing and feature extraction. To construct our summaries, we applied shot boundary detection and multiple keyframe extraction to locate important segments within the footage along with keyframes, which could be used to accurately represent detected shots within the final summary. This information, in combination with face detection and motion estimation, was used to determine the *importance* of each shot and to select only a sufficient number as would fit in the 4% target duration. Our resulting summaries make use of usability considerations such as gradual transitions and appropriate audio tracks, as well as an information overlay.

Despite our human factors approach to summary construction and presentation, we performed more poorly than we hoped, particularly in the ease of use scoring. As a result, we were motivated to undertake detailed failure analysis and investigate the possible reasons for our poor performance. The below sections outline the summarisation approach taken, followed by an analysis of the results provided by the NIST evaluation, our failure analysis and recommendations for improvement of our approach, as well as considerations for the evaluation procedure.

TVS' 07, September 28, 2007, Augsburg, Bavaria, Germany.

2. SUMMARISATION APPROACH
The following sections outline the methods used to construct our final summaries.

2.1 Video Processing
Our summarisation approach is a 2-stage process. In the first stage a suite of video analysis techniques is applied as a pre-processing stage prior to summary construction. This extracted data describes features of the original footage which is later used to create the summaries. For each original video, the processing provides a set of shots each containing one or more keyframes. Each keyframe is accompanied with the following information: the keyframe's frame number, the frame number for the beginning of the keyframe's shot, the frame number for the end of the keyframe's shot, a measure of the average motion within all frames in the keyframe's shot, the number of faces detected in the keyframe and the bounding rectangle for each face. This is described in more detail below.

2.1.1 Shot boundary detection & keyframe selection
Our shot boundary detection and keyframe selection processes occur simultaneously allowing the extraction of multiple keyframes per shot. The shot cut detection method works in the

uncompressed domain based on a histogram comparison method. Each shot of the video is analysed to compute an activity measure from which the number of keyframes per shot is determined. The larger the amount of activity in a shot, the more keyframes allocated to that shot. Keyframes are identified using a polygonal approximation of the activity measure graph, which is computed using RGB histograms. Once this process is complete, the frame numbers marking the beginning and end of each shot are output, along with the keyframes (saved as JPEG image files). One limitation of our detection methods is that the number of keyframes to be extracted per video must be defined in advance.

2.1.2 Motion Activity Detection

This three-stage process, achieved using the AceToolbox [8], allows us to describe the level of motion activity in each shot. First, the MPEG-1 motion vector data is extracted from the video. Next, the percentage of non-zero blocks in the frame (where a high percentage indicates higher motion activity) is calculated for each frame in the video. Finally, this per-frame data is used along with the shot-boundary data calculated previously to compute an average motion measure for the entire shot. As a result, each keyframe in a given shot will be assigned the same measure of motion activity. We favour this over the provision of a per-keyframe measure of motion activity, in order to preserve the contribution of all frames in a shot, and not just the given keyframes. By only providing this data at the keyframe level, the overall motion activity might be inaccurately portrayed and affect the ranking of shots within the generated summary

2.1.3 Face Detection

Our face detection processing extends the Bayesian Discriminating Feature (BDF) originally proposed by Liu [6] for detecting frontal faces in grayscale images. Using a statistical skin colour model [1], we can detect multiple faces at various sizes and orientations within colour images. Ideally this processing would be carried out for each frame of the original footage, however, for efficiency we only perform this operation on the detected keyframes. While this potentially results in the loss of information, such as the prevalence of faces across shots, keyframe only processing ensures efficient processing while still providing enough information to reliably enhance the summary construction.

2.2 Summary Construction and Presentation

The keyframe-based summaries are constructed by automatically generating Extensible MPEG-4 Textual (XMT) documents [4], using feature information extracted during the video processing. These documents are then converted to MPEG-1 content. Below is an outline of the major steps in our summary construction.

2.2.1 Selection of Representative Keyframe

In our summaries, each shot is represented by a single keyframe and the included shots are displayed in temporal order. Often, there are more shots within the original footage than can fit into the target duration (4% of the original video duration). In order to select the most important shots for inclusion, we use a simple weighting to rank them, in which the most important factor for inclusion is the *relative duration* of the shot (see below).

Rank Score for Shot = (Number of faces/Maximum Faces in Footage x 0.2) + (Amount of motion x 0.2) + (Number of keyframes/Total Keyframes x 0.6)

Each shot is displayed for a short period (minimum of 1 second to a maximum of 2.5 seconds) depending on its features. The minimum of 1 second provides the human visual system (HVS) adequate time to attend to a new image and visually process it [3]. The greater the number of objects in the scene, the more time the HVS will require to process it but providing this additional time has also been shown to increase the performance and accuracy of its later recall [3]. For these reasons we decided to assign keyframes extra time on screen if they met certain criteria. The extra time was informed by experiments, such as [3], examining the relationship between the number of objects present in a scene, the time taken to visually attend to that scene and the subsequent recall of the objects. For each face detected, an additional 0.2 seconds is allocated to the in-summary display. Similarly, if the shot contains a high degree of motion, indicative of a lot of activity, it is assigned an extra 0.5 seconds on screen.

Once ranked and the required display duration calculated, we then select the top *N* shots that fit within the target duration. Finally, an overview keyframe for the shot is then selected from the group of representative keyframes previously detected for that shot. As faces are attracters of overt visual attention [12,7], keyframes known to have faces present are preferred for inclusion, alternatively the mid-frame from the group is selected.

2.2.2 Overcoming Change Blindness.

It has been shown in numerous studies [10], that people are extremely bad at detecting change within a visual scene. Surprisingly large changes can be made in footage without a viewer noticing e.g, the addition or removal of objects; the substitution of one person for another. This phenomenon, known as "*change blindness*", is particularly worrisome for a keyframe-based video summary of rushes where a number of scenes will be rapidly presented in quick succession. A common cause of change blindness in video is the use of hard cuts between scenes and this explains why continuity errors in films go unnoticed by many viewers [5]. Change detection is improved with visual attention and time, and has been shown to take up to 1,500 ms in approximately 41% of people [2]. In order to promote change detection in our video summary, we ensure that the keyframe is displayed for a minimum of 1,000 ms and we also provide cues to the user, notifying them of impending shot changes. Instead of hard cuts, our summaries use a short cross-fade transition (500 ms in length) between keyframes to visually indicate the shot change and overcome the change blindness issue. An audio cue played just before a shot cut is a further cue to an upcoming change.

Figure 1. An example of the composed summary showing the amount of motion, number of people and offset of this section of the summary within the original video

2.2.3 Audio

The inclusion of good quality audio in video content, particularly in summary video, is known to enhance its perceived usability and reduce the ability of a viewer to detect impairments in the video [11]. Based on this we overlay each summary with one of 10 randomly selected ambient audio tracks. Each track is approximately 30 seconds in length but is looped if required. Audio is also used as a mechanism to overcome change blindness. A short "beep" is also added prior to shot changes to provide an auditory cue to help overcome "change blindness".

2.2.4 Information Overlay

Due to the nature of a keyframe based approach to video summarisation, some contextual information on the activities within the scene are lost. To help overcome this, an information overlay is included at the bottom of the generated summary (see Fig 1). The overlay is low contrast so as not to draw overt attention from the displayed keyframe and it does not occlude any of the keyframe's detail. The overlay is updated when a new shot is displayed on screen and indicates the amount of motion and number of people for that shot. The timeline also illustrates what portion of the original footage the current keyframe covers.

3. RESULTS AND DISCUSSION

Overall our summarisation submission performed poorly, with a few minor exceptions. The full description of the results can be found in the Workshop overview paper [9].

While our approach reported extremely compressed summaries relative to other groups (with most falling significantly under the target duration [9]), this however appears to have been to the detriment of the summary coverage (or the extent to which the summary deals with the major points). The inclusion results place DCU (mean: 0.38; median: 0.38) among the 5 lowest scoring participants. The possible reasons for our poor inclusion scores are detailed in the Failure Analysis sections below.

We were particularly disappointed with our low '*ease of use*' scores (mean: 2.53; median: 2.67) which placed us second worst out of 25 participants, especially as we placed heavy emphasis on human factors considerations in our summarisation approach. Interestingly, we scored particularly well for total time taken for judgment. Our summaries were among the quickest to use based on the median total time and non-paused time for judging inclusions (mean: 71.44; median: 70.83). This however, seems to be at odds with our ease of use scores. It could be expected that the more usable a summary is, the easier it should be to locate inclusions bringing the 'ease of use' scores and judgment times into alignment, however, there is no indication of such a correlation. This may indicate that the 'ease of use' scores are particularly sensitive to the content contained within the summary or that 'ease of use' in fact reflects coverage as opposed to flow. Despite little effort to remove redundant and duplicated content, our submissions scored on par with other groups (mean 3.67; median: 3.67).

4. FAILURE ANALYSIS

Our unexpectedly poor performance in the summarisation task encouraged us to undertake detailed failure analysis. First, as our approach to selecting content and dealing with coverage was simplistic, we examined deficiencies in the construction of our summaries, which may have resulted in our low inclusion scores. Second, as the focus of our effort was in human factors of summary construction and presentation and we were surprised by our low 'ease of use' scores, consequently we investigated the evaluation methodology employed by NIST to determine if there were any possible 'teething issues', which may be attributable to our lower than anticipated results. As part of this investigation we explored possible refinement and following this conducted re-evaluations of our summaries. We outline in detail below our methods and the outcomes relevant to both our submission and the overall evaluation approach adopted.

4.1 ANALYSIS OF NIST EVALUATION

During our failure analysis we noted some unexpectedly low inclusion scores for a number of our summaries. In combination with our surprisingly low usability scores, this motivated us to repeat the NIST evaluation in-house with our own users. Some specific aspects of the methodology, based on observation analysis of video playback and evaluator comments from NIST also raised additional concerns that we wanted to investigate. Only allowing an evaluator to view the summary once (pausing as required) does not appear to be an ecologically valid approach to review of summary content. In real world use a viewer can rewind, pause, play, fast-forward and skip through content as they desire enabling them to return to a specific segment should they miss an item, for example. We believe this artificial review may have contributed to lower inclusion scores. By not allowing evaluators to return to investigate a segment, should they miss and item or have doubt about the inclusion, it is likely that that item would not be deemed to have been included in the summary and could cause inaccurate coverage to be reported for the summaries. While this might only apply to a keyframe-based approach it may also more generally impact other participants' approaches.

To probe these concerns, 5 participants (all students in the School of Computing at DCU) were opportunistically selected to review the 13 worst performing DCU summaries under the guidelines provided by NIST. This investigation used the original summaries submitted for evaluation by NIST without any alterations. Initially the evaluation mirrored the NIST methodology with participants being instructed to play the summary just once, then complete the inclusion ratings and usability questions. They were next allowed to freely review the summary a second time, viewing it in any manner they desired (including skipping forwards and backwards or replaying the summary as often as needed). The results of this re-evaluation can be found in Table 1.

A significant increase in the ease of use scores was reported in the second evaluation, from 2.17 to 3.06. However, this may not be indicative of issues with the NIST evaluation (see 4.1.1) as it may be a result of biasing in favour of our submission as a result of participant association with DCU. The duplication scores overall remained consistent with the reported scores from NIST.

The inclusion scores roughly correlated with the NIST reported values on the first pass, with a minor but acceptable increase. The increased inclusion scores may be a result of slightly less stringent evaluation than was conducted by NIST evaluators. Interestingly however, by allowing a free review of the content a 9% average (8% median) increase in the reported inclusions was recorded for summaries.

The results of this evaluation indicate that a free review, which may more ecologically valid, would lead to an increase in the reported coverage. While, the NIST adopted approach highlights the ability of the reviewer to accurately extract an understanding

Table 1. Comparison between NIST and Internal Evaluation Results. Additional Increase for inclusions after free review is contained in the brackets but not included in the score for comparative purposes.

	NIST			DCU NIST Comparison			Summarisation Changes		
Video	**Inclusion**	**Ease**	**Dupl.**	**Inclusion**	**Ease**	**Dupl.**	**Inclusion**	**Ease**	**Dupl.**
MRS035132	0.17	3.67	5.00	0.17 (+0.03)	3.00	3.60	0.33 (+0.11)	4.67	1.67
MRS042543	0.22	2.33	3.00	0.32 (+0.08)	2.80	3.40	0.39 (+0.03)	4.00	3.33
MRS043400	0.18	2.67	3.00	0.36 (+0.04)	2.60	1.80	0.33 (+0.11)	3.33	2.33
MRS044500	0.22	3.00	4.00	0.32 (+0.07)	3.20	3.20	0.42 (+0.08)	3.33	3.00
MRS044731	0.11	2.67	3.00	0.27 (+0.08)	2.60	3.40	0.50 (+0.00)	4.00	3.33
MRS048086	0.08	1.00	4.33	0.17 (+0.10)	2.80	3.80	0.36 (+0.08)	3.67	3.00
MRS145918	0.39	1.67	3.67	0.38 (+0.12)	3.40	4.20	0.47 (+0.03)	4.00	3.67
MRS155017	0.20	2.00	3.67	0.28 (+0.03)	2.40	3.00	0.61 (+0.06)	3.67	3.67
MRS155534	0.25	2.00	3.33	0.45 (+0.13)	3.00	2.60	0.56 (+0.08)	2.67	2.67
MRS157444	0.08	1.00	2.67	0.27 (+0.17)	3.60	4.40	0.58 (+0.03)	4.00	1.00
MRS157475	0.22	3.00	4.67	0.37 (+0.07)	4.20	4.20	0.56 (+0.06)	4.67	2.67
MRS158385	0.53	2.00	4.67	0.43 (+0.07)	3.40	4.40	0.67 (+0.06)	5.00	4.00
MRS336905	0.31	1.67	5.00	0.30 (+0.12)	2.80	4.80	0.36 (+0.06)	3.33	3.67
Mean	**0.25**	**2.17**	**3.79**	**0.32 (+0.09)**	**3.06**	**3.60**	**0.47 (+0.06)**	**3.87**	**2.92**
Median	**0.22**	**2.00**	**3.67**	**0.32 (+0.08)**	**3.00**	**3.60**	**0.47 (+0.06)**	**4.00**	**3.00**

of the content of the original from a single viewing, a free review could more accurately interrogate the overall coverage of a summary. There is clearly merit to both single constrained and detailed free reviews, and it is another factor that can be considered when interpreting the summary evaluation.

4.1.1 Additional Comments

Audio: It was noted in the Assessors Comments provided by NIST [9] that at least one, if not more, assessors opted not to use the headphones as they found them to be a distraction, particularly when reviewing the high speed original footage. As audio may not have had been reviewed by all evaluators we chose to remove audio playback from our own evaluations. Additionally, we are now unsure if the inclusion of audio in our summaries contributed to our performance in the task.

Groundtruth Validity: During our evaluations, several of our participants commented on issues with the ground-truth dataset. Despite the guidelines for groundtruth construction, provided by NIST [9], indicating that insubstantial content should not be included we noted a few such examples existing within the groundtruths. For example, MRS157475 contained "woman fixes clapper board", despite the groundtruth instructions indicating that content such as clapperboards *should not* be included. Additionally, some groundtruth items were extremely difficult to locate in the original footage and were not representative of the overall content to be summarised. An example of this is in MRS157444, where the "head of young man enters from left of shot in front of bearded man," which represents only 6 seconds from the 35 minutes of original footage. While the inclusion of this content increases the complexity of the summarisation task, it is also important to consider that this may also have negatively impacted resulting inclusion scores for certain summaries.

4.2 ANALYSIS OF OUR APPROACH

Following the release of the evaluation results, we reviewed our worst performing summaries to expose the most likely factors which detrimentally impacted our 'inclusions' and 'ease of use' scores. We determined the two likely issues to be:

Forsaking coverage for brevity: On average our summaries were over 17 seconds shorter than the maximum target length (4% of the original). This was caused by the crude means in which we selected keyframes for inclusion in the summaries. In our summaries, each shot is represented by a single keyframe, and the number of shots detected in the original content is lower than the total time required to display them. Thus, our approach returned summaries with a large amount of unused space. These summaries tend to contain less useful content and an excess of short uninteresting and non-representative shots. Furthermore, by providing only a single keyframe per shot, there often lacked enough contextual cues to accurately determine some groundtruth inclusion items, particularly if a groundtruth item related to movement of characters in a scene or camera activity. A simple but effective solution to this would be to redistribute the extra space and assign additional keyframes to some or all shots in the summary, depending on the amount of free space.

Inclusion of redundant and confusing content: A large number of our summaries contained redundant content such as blank frames, test-cards, clapper-boards, etc. The presence of this type of content often negatively impacted on the flow and usability of the summary and created an impression of duplication. Where clapper-boards were present, they often obscured content in the scene. This probably made determining if an item was included more difficult for an evaluator. More importantly, the presence of this redundant content in the summary means that more relevant or representative keyframes were not included and this is likely to have contributed to our poor inclusion performance. Potentially the majority of the redundant content could easily be removed by automatic processes, for example, trained to look for the patterns of a test-card or an almost blank frame.

We decided to investigate the effect of these problems in our summary by making modifications to our approach and re-testing revised summaries under the evaluation guidelines provided by NIST [9]. We selected the 13 worst performing summaries in the inclusion and ease of use category (see Table 1) and generated revised summaries for these, which addressed the two major issues identified above. As there was insufficient time to create an automatic process to remove redundant content, we simulated this automatic extraction by manually removing content that we deemed an automated system could easily identify. We also revised our keyframe selection approach to allow for redistribution of free space in the summaries to the display of additional keyframes for the more important shots, as follows. After investigation, we found that high motion shots are most

likely to contain camera activity such as pan or zooming and the high motion is in itself an indicator of requirement of additional contextual cues, i.e. from a single keyframe we may not be able to tell in which direction a person enters or leaves a shot from, but with an additional frame more information on their movement would be available. So, first, any shot with a high motion activity is assigned an additional keyframe, if one can be accommodated. We then use our previous importance calculation to rank the shots and assign additional keyframes in sequence until there is no more remaining space in the summary. The mid-keyframe is always chosen and the subsequent frames are chosen to the left and right of the mid-frame to give maximum coverage of the shot. Keyframes close to the start and end of the shot are eliminated from inclusion to further reduce the possible inclusion clapper-boards and blank frames.

To determine if our assumptions regarding our poor performance were correct, we conducted an evaluation on the modified summarisation approach. For this, 3 students from within the school were opportunistically selected. Previous participants were not involved in this investigation to mitigate against exposure issues. Each participant was presented with the new versions of the 13 worst performing summaries using an identical method to the previous evaluation outlined in 4.1.

A significant improvement in the coverage scores was achieved. The median inclusions doubled in our new approach when compared with the original NIST evaluated summaries. This indicates that the brevity of our summaries did impact of the overall coverage of the summary. It also suggests that the addition of multiple keyframe per shot provided contextual cues that allowed evaluators to more easily determine the presence or absence of groundtruth items.

If we assume that participants were somewhat biased towards the DCU submission in the NIST comparison evaluation, based on a significant increase over both the NIST baseline and DCU NIST comparison evaluation we can reliably determine that our modifications resulted in a large increase in the summaries' 'ease of use'. This signals that the use of multiple keyframes combined with removal of redundant content allows for the creation of a summary with better overall flow and improves the viewers understanding of the original footage's content. The inclusion of extra keyframes, however, negatively impacted on the duplication scores, with participants noting an increase in the duplication of content within the new summaries. With the inclusion of more keyframes per shot, more repetitious, retake content from the Rushes is included, accounting for the negative decline in performance in our modified summaries.

These results are extremely encouraging and demonstrate a major improvement in both inclusion and usability for our modified summaries. The evaluation has also confirmed our hypotheses regarding the deficiencies in our original summaries and the modifications made to the summary construction provides a more complete summary without sacrificing human factors considerations.

5. CONCLUSION

Our keyframe based approach to video summarisation brings together video processing techniques and human factors considerations. After reviewing the results from the NIST evaluation, we carefully considered features of our summaries that negatively impacted our performance. This allowed us to further refine our summarisation approach. The inclusion of multiple keyframes per shot and the removal of redundant content, such as clapper-boards were demonstrated to be effective improvements. While the automatic removal of redundant content has yet to be implemented, it clearly merits attention and we will pursue this as part of our future summarisation efforts. Finally, as part of our failure analysis we also explored the NIST evaluation framework used and highlighted an alternative approach to summary evaluation using free unguided review.

ACKNOWLEDGMENTS

This work was partially supported by the Irish Research Council for Science Engineering and Technology, Science Foundation Ireland under grant 03/IN.3/I361 and by the European Commission under contract FP6-027026 (K-Space). We are grateful to the AceMedia project (FP6-001765) which provided us with output from the AceToolbox image analysis tooklit.

REFERENCES

[1] Cooray, S. and O'Connor, N. A Hybrid Technique for Face Detection in Color Images. In *IEEE Conf. on Advanced Video Surveillance (AVSS'05)*, Italy, Sept 15-16, 2005

[2] Hollingworth, A., Williams, C.C. and Henderson, J.M. To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes, *Psychonomic Bulletin & Review*, 2001, 8, 4, 761-768

[3] Irwin D.E., Zelinsky G.J. Eye Movements and Scene Perception: Memory for Things Observed, *Perception & Psychophysics*, 2002, 64, 6, 882-895

[4] Kim, M., Wood, S. and Cheok, L. Extensible MPEG-4 textual format (XMT), Proceedings of *the 2000 ACM workshops on Multimedia*, Los Angeles, US ACM Press, New York, NY, USA, 2000, 71 - 74

[5] Levin, D.T. and Simons, D.J., Failure to detect changes to attended objects in motion pictures, *Psychonomic Bulletin and Review*, 1997, 4, 501-506

[6] Liu, C. A Bayesian discriminating features method for face detection. *IEEE Tran. on PAMI*, 25:725-740, June 2003

[7] Neilsen, J. Talking-Head Video Is Boring Online, AlertBox, Dec 2005, retrieved from: useit.com/alertbox/video.html

[8] O'Connor, N., Cooke, E., le Borgne, H., Blighe, M. and Adamek, T. The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification, In *Proceedings 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005, London, U.K.

[9] Over, P., Smeaton, A.F. and Kelly, P. The TRECVID 2007 BBC rushes summarisation evaluation pilot. In *Proceedings of the TRECVID Workshop on Video Summarisation (TVS'07)*, Augsburg, Germany, September 28, 2007, ACM Press, New York, NY, 2007, 1-15.

[10] Rensink, R.A. and Clark, J.J. To See or Not to See: The Need for Attention to Perceive Changes in Scenes, *Journal of Psychological Science*, 1997, 8, 5, 368

[11] Winkler, S. Issues in Vision Modeling for Perceptual Video Quality Assessment, *Signal Processing*, 1999, 78, 2,231–252

[12] Yarbus, A. *Eye Movements and Vision*, Plenum Press, 1967.