映像による個人行動記録・要約システムとその評価
— 注目シーン検出と要約の評価 —

大出 純哉　中村 裕一　大田 友一
筑波大学 機能工学系
(yuichi@esys.tsukuba.ac.jp)

Attention-based Experience Spotting and Its Evaluation

Jun'ya OHDE, Yuichi NAKAMURA, Yuichi OHTA
Institute of Engineering Mechanics and Systems,
University of Tsukuba

あらまし: 本研究では、人間が身体装着したカメラを用いて獲得した日常行動映像を構造化、要約する手法を提案してきた。この手法では、利用者が何かに注目したシーンを映像中から抽出するために、画像中の動きを求め、得られた動きパラメータを用いて注目対象領域を検出する。本稿では、本手法の妥当性を検討することを目的とし、新しい評価方法とその結果について述べる。

1　はじめに

個人の見聞きした情報を映像を用いて記録し、必要に応じてそれを再現するようなシステムには、記憶の補助、記憶の共有等の面で多くの用途が考えられる。計算機環境の進歩にともない、実際にこのような目的で研究が始められている [1][2][3]。

我々はこのような研究の一つとして、比較的粒度の細かい行動記録・要約を実現する手法を提案してきた [4][5]。ここでは、“何をしたのか” だけではなく、“どのようにしたか” が分かる記録を得ることが目的となっている。

そのためには、映像を逐一記録することが必要になるが、得られるデータ量が膨大になり、かつ、内容は冗長なものとなる。そのため、ただ単に映像を記録しただけでは、利用者が映像を見直す際に多大な労力を必要とする。 我々はこの問題を解決するために、利用者が何かに注意を引かれ、それを注目した場面を検出することによって、映像の要約、構造化を行なう方法を提案してきた。

これまでの研究により、概ね良好な結果が得られることがわかってきたが、ここで目的とする構造化や要約には大きな自由度があるため、得られる結果に対する評価、及び、手法の妥当性を検討することが必要である。そのため、本研究では、我々の提案する動き推定処理、要約方法の各々について評価を行い、その有効性について考察した。

以下、本稿では、まず我々の提案手法について簡単に説明し、次にその評価方法とその結果について詳しく述べる。

2　注目シーンによる映像要約

2.1　注目動作と注目シーン

カメラを頭部に装着し、装着者の視点から撮影することによって、常に装着者の目前にあるものを撮影することができる。これにより、装着者の記憶と映像として得られたデータとの照合がとり易くなることが期待できる。

我々は、これにより得られた冗長なデータを解析するために、装着者が何かに注目して見ようとしたシーン、すなわち、最もよく覚えているであろうシーンを検出、提示することを提案した。これを注目シーンと呼ぶ。そのために、利用者の注目動作とそれに伴う画像特徴を定義してきた。

それには、以下のような 2 つがある。

対象を注視する動作:　人は何かに興味を持つと、その対象をじっと見続けようとする。見続けようとする典型的な動作としては、図 1(a) のように人が静止したまま静止している対象物をじっと見る場合や、図 1(b) のように動いている物体を首を振りながら、または体を動かしながら視界に入れ続けるよう追いかける、などといった動作がある。

長時間の視点の停留:　デスクワーク、他人との会話、講演の聴講のような場合には、同じ場所で比較的長時間、同じ出来事を見続ける。この場合は、図 1(a) に体の揺れが複合的に加わったものになる。この場合、シーンの長さが長時間におよぶ。

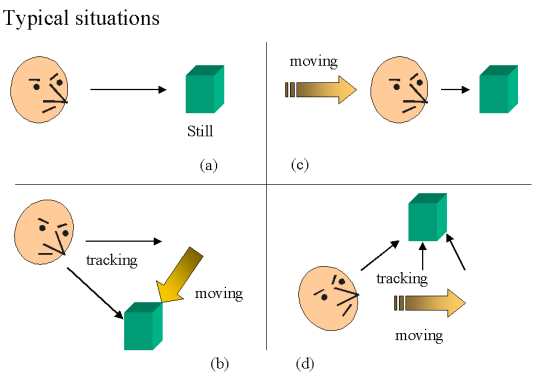


図 1: 注目時の動作

本研究では、上記の 2 つのシーンをそれぞれ積極的注目シーン、停留シーンと呼ぶ。

2.2 注目シーンと画像特徴

前節で述べたような注目動作を画像処理によって検出するためには、頭の動きによる画像中の見かけの動きと注目対象の動きを推定し、必要に応じて分離しなければならない。

ここで、視点の停留を検出する場合には、画像中全体に渡って見かけの動きが小さいことを検出すればよい。また、注目動作を検出する場合には、図 2 に示すような画像中の 2 つの動きを分離する必要がある。ただし、図 2 右のように、注目対象が複雑な動きをする場合もある。

いずれの場合にも重要なことは、注目対象が画像中に停留することである。つまり、比較的短時間 (例えば 30 フレーム程度) ではあるが、注目対象をじっと見続けるため、画像内の特定の部分またはその近傍に上記の領域がとどまることである。

本研究ではこの点に着目し、以下で述べるような注目シーン検出処理を行う。

3 注目シーンの検出

注目シーンの検出は以下のような手順で行う。

1. まず、頭の動きに起因する見かけの動きを推定するために、時間的に 1∼ 数フレーム離れた 2 枚の画像間の対応関係を求める。
2. 求まった対応関係によって、視野の移動が小さいと判定される場合には、隣接するフレームをまとめて “停留シーン” とする。
3. 停留シーンと判定されなかったフレームについて、注目対象の検出を行う。注目対象が存在し、その見かけの位置が大きく変化しない部分を、 “積極的注目シーン” とする。

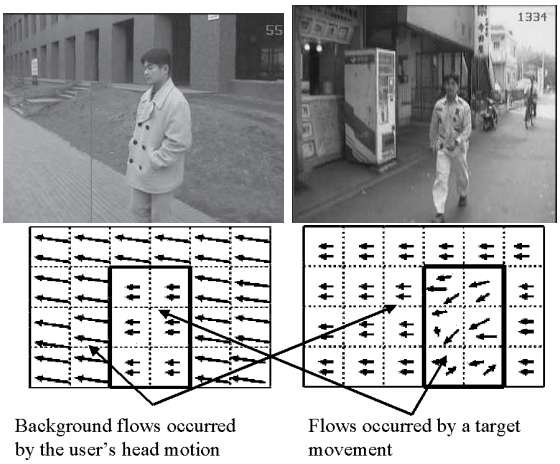


図 2: 積極的な注目と画像特徴: この例では、歩いている人に注目しているとする

以下でその概略を説明する。

3.1 画像間の動きベクトル推定

個人行動映像には、目前から数メートル先までに種々の物体が存在し、被写体を同一平面として近似できる場面が少ないため、中心投影のモデルを用いる。

これは、画像内の点 $\mathbf{x}$ の動き $u(\mathbf{x})$ は、カメラの並進運動 $\mathbf{t} = (t_1, t_2, t_3)^T$, 回転運動 $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)^T$ を用いて、以下の式で表される。

$$u(\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\mathbf{A}\mathbf{t} + \mathbf{B}\omega \qquad (1)$$

ただし、

$$\mathbf{A} = \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} (xy)/f & -(f^2+x^2)/f & y \\ -(f^2+x^2)/f & -(xy)/f & -x \end{bmatrix}$$

また、f は焦点距離、$(x, y, Z(\mathbf{x}))$ は中心投影の座標系における点 $\mathbf{x}$ の 3 次元位置である。

ここで、座標値 $\mathbf{x}$ における輝度値を $I(\mathbf{x})$ とすると、以下のエラー E が最小になるように上記パラメータを決定する。

$$E = \sum_{x,y}(I(\mathbf{x}, T) - I(\mathbf{x} - \mathbf{u}, T - \delta t))^2 \qquad (2)$$

処理手順は次のようになる。

1. 上記のモデルは、平行移動成分が小さい場合に不安定になる。そのため、動き推定を行う前に、静止区間 (動きがない、又は非常に小さい区間) を動き推定処理の対象から外す。静止区間の検出は、ガウシアンフィルタをかけた 2 枚の画像の差分を計

算し、差分が閾値よりも小さいことを条件とする。
2. 平滑化とダウンサンプリングによる多重解像度画像の生成。ここでは、入力画像の 1 辺に対して, $1/2$, $1/4$ の大きさの画像を生成する。以下を最も高い解像度まで順に行う。
3. カメラ運動パラメータの初期値を付与する。ここでは、一つ粗い解像度で推定された動きパラメータを初期値とする。ただし、最も粗い解像度の初期値は、1 つ前のフレームに対する動きパラメータとする。
4. 式 (2) のエラーが最小値になるパラメータを繰り返し計算により求める。本研究では、Levenberg-Marquardt 法を用いる。

3.2 動き推定の安定化

頭部の動きが急激に変化する場合、テクスチャが少ない場合、また、注目対象が大きいために頭の運動による見かけの動きとの分離が難しい場合には、前節の計算が不安定になることがある。

そのため、画像中の点の動きベクトルの大きさを拘束条件として加えることによって安定化を計る。短時間 ($1 \sim$ 数フレーム) では、カメラ移動が小さいと仮定できるためである。

まず、式 (2) に対して、安定化項 $g_i(\mathbf{t}, \omega)$ を加えた式を考える。

$$E_a = E + \lambda \sum_i g_i(\mathbf{t}, \omega)^2 \tag{3}$$

ここで、$g_i(\mathbf{t}, \omega)$ はパラメータ $(\mathbf{t}, \omega)$ で微分可能であるような関数から選ぶ。例えば以下のように定義する。

$$g_i(\mathbf{t}, \omega) = \exp^{a_i/K} \tag{4}$$

ここで、a_i は拘束、K は拘束の強さを調整するパラメータとする。

次に、式 (4) において、

$$a_i = d(\mathbf{x_i}) - d_{max}(\mathbf{x_i})$$

とする。ただし、$d(\mathbf{x_i})$ はある点 $\mathbf{x_i}$ の動きベクトルの大きさとし、$d_{max}(\mathbf{x_i})$ には、理想的な状況で考えられる動きベクトルの上限値付近の値を与える。

本研究の実験では、点 $\mathbf{x_i}$ として 3 点を選び、3 点の移動ベクトルの大きさを拘束条件としている。

3.3 停留シーンの検出

3.1 節で述べた方法により、静止区間を検出する。

次に、前述の方法で求まった動きベクトルを用いて、静止区間どうしを同一シーンとして統合する。そのために、2 つの静止区間にはさまれた区間で、3.1 節で求められた画像中心の見かけの動きベクトルを足し合わせていく。その結果、動きが閾値よりも小さい時、同一シーンとして両者をまとめる。

3.4 注目シーンと注目対象の検出

注目シーンの検出は、見かけの動きを頭の動きに起因するものと注目対象の動きに起因するものに分離することにより行う。

まず、上記の方法で求まった動きパラメータと式 (1) を用いて、時刻 $T - \delta t$ の画像 $I_{\mathrm{T}-\delta \mathrm{t}}$ を時刻 T のカメラ位置からの見え方 $I^T_{\mathrm{T}-\delta \mathrm{t}}$ に変換する。

次に、この変換画像と時刻 T の画像 I_{T} の間で、小ブロックどうしの相関をとり、相関の低い領域をまとめることで注目対象の候補として抽出する。

この領域がある時間以上に渡って、同様の場所に存在する時、このシーンを注目シーンとして検出する。

紙面の都合上、詳細については [5] にゆずる。

4 システムの評価方法

本研究では、システムの評価として、注目シーン検出に対する物理的な評価と生成した要約に対する人間にとっての良さの両方を評価する。

4.1 注目シーン検出に対する評価

注目シーン検出の最も重要な部分となる、画像間の動き推定処理の評価を行った。注目対象抽出処理部分の評価は今後の課題となっている。

評価方法は以下のようになる。

1. 画像 I_i と奥行き画像 D_j の組を用意する。実験では、普段対象となることの多い室内シーンのテクスチャを I_i とし、室内、廊下、近接物体があるシーン等を簡略化した奥行きを D_j として選んだ。
2. 正解となる動きパラメータのセット $(\mathbf{P}_k)$ をあらかじめ多数用意する。
3. D_j を仮定し、$\mathbf{P}_k$ によって I_i を変換することによって、新たな見え方画像 $I^{'}_{ijk}$ を生成する。また、この際に画像内の各点の移動ベクトル $\mathbf{v_a}(\mathbf{x})$ を計算し正解データとする。
4. 既存の画像 I_i と新たな見え方画像 $I^{'}_{ijk}$ 間の動き推定を行ない、これによって推定された動きベクトル $\mathbf{v_b}(\mathbf{x})$ と正解データ $\mathbf{v_a}(\mathbf{x})$ を比較することによって、動き推定の評価とする。

4.2 要約に対する評価

注目シーン検出による映像の要約がどれだけ役に立つか、また、個人の考える要約に対する近さを検証することを目的とする。

従来の研究では、映像要約の良さを評価する物理的、心理的な尺度として、映像に関連する問題に答える際の、所用時間、正答数等が用いられてきた。Smithらは、Video Skimmingの良さを評価するために、映像の内容に関する問題に解答するための時間を用いている[8]。また、Eldridgeら[9]や川嶋らの研究では、問題に正答した数により評価を試みている。

本研究においても、問題解答の所用時間を一つの評価基準とする。しかし、これらには、本人がその要約に満足できるかどうかという点での評価が欠けている。つまり、本人にとって重要であった出来事が網羅的に捉えられているかどうかを評価する必要がある。

そこで、各々の被験者に要約に必要なシーンを選択してもらい、選ばれた結果との一致度をもう一つの評価基準とする。

解答時間による評価

まず、被験者を2つのグループに分ける。一方のグループは本システムで作成した注目シーンによる要約を、他方のグループはビデオデッキを利用して映像を閲覧する。各々のグループに対して、映像内容に対する4つの質問に解答してもらい、すべての問題に正解するのに要する時間を計測した。質問の内容は全く同じものとするが、実験は各被験者に対して個別に行い、相談することはできない。

要約としての主観的評価

以下のような方法をとる。

1. 映像から複数の方法で多数のフレーム(画像)を選び出す。ただし、見た目がほぼ同じフレームがある場合には、それらのフレームを同一のものと見なす[1]。
2. 1で集めたフレームの集まりから、各人に指定枚数(N)のフレームを選び出して貰う。つまり、被験者に自分が最も良いと思う要約を作るための材料を選んで貰う。

次に、ある方法による要約結果(フレーム集合 F_i)と人為的に生成された要約(フレーム集合 H_j)の一致する割合を各人への合致度と呼び、以下のように定義する。

$$\text{合致度}_{ij} = \frac{F_i \cap H_j\text{に含まれる枚数}}{N}$$

ただし、人に生成される要約には個人差が大きく現れるため、その点を考慮することが必要となる。本研究では、被験者集団により各々のフレームが選ばれた回数を考え、その回数が多いものを一定数(N)集めたものを 平均的要約(H^*)と考える。これを個人差を考えない場合の最良の選択と考える。

また、各人の要約と平均的な要約に対する合致度に対して、それぞれの要約方法がどの程度近付いているかを評価する。本研究では、この割合を到達率と呼び、以下のように定義する。

$$\text{到達率}_{ij} = \frac{F_i\text{と}\ H_j\text{の合致度}}{H^*\text{と}\ H_j\text{の合致度}}$$

このようにして合致率を計算することにより、注目シーン検出による映像要約の有効性が検証できる。

5 実験

実際の映像に対し、これまで述べてきた処理とその評価を行った結果を示し、注目シーンの検出の有効性を検討する。

5.1 動き推定の評価

4章で説明した動き推定の評価を行った. 図3に実験で用いた奥行き画像の例を示す。図3(a)は通常の室内シーン、(b)は廊下、(c)は近傍に大きな物体がある場合を想定している。

動きが小さい場合として、平行移動成分を $t_i = 0, \pm 50(mm)$、回転移動成分を $\omega_i = 0, \pm 3(\text{degree})$ として画像を生成した。これによって得られる各点の移動ベクトルの大きさは、$0 \sim 20$ 画素程度である。

また、動きが大きい場合として、平行移動成分を $t_i = \pm 100(mm)$、回転移動成分を $\omega_i = \pm 5(\text{degree})$ として画像を生成した。各点の移動ベクトルの大きさは $30 \sim 50$ 画素程度である。

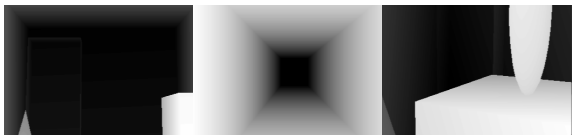
これらの生成画像に対して、動き推定を行って、正解の動きベクトルとの比較を行った結果を図6に示す。拘束に関しては、紙面の都合上、λ, K の選び方のうち、良い振る舞いを示したものだけをあげる。

結果を見ると、動きが小さい場合には、拘束の有無に関わらず、誤差1画素程度の精度が得られることが分かる。これは、注目対象を検出するためには十分な精度であると言える。

動きが大きい場合には、平均的な誤差が悪くなり、後の注目対象検出処理に影響が出てくる。拘束をかけることによって、平均誤差は少し良くなり、得られる動きパラメータも図5に示すように、良いものが得られる場合がある。しかし、平均誤差は大幅には良くならない。この点については今後の対策が必要である。

ここで重要なのは、拘束をかけることによって、極端に大きな動きベクトルが出てくるのを防ぐ効果が現れていることである。拘束なしの場合には、移動ベクトルの推定値が大きくなることによって、停留シーン

[1] 各々の方法によって抽出された画像同士が似ている場合には、ほぼ同等の意味を持つ画像の中から被験者が一枚を選択することになる。このような場合、特に根拠なく選ばれるため、システムの正当な評価を妨げることになる。ただし、見た目の同一性は被験者により判定してもらう。



(a) depth1　(b) depth2　(c) depth3

図 3: 動き推定のための奥行き画像: (a) は典型的な室内シーン、(b) は廊下等、(c) は視野の中に，大きな物体が近接して入っている場合を想定している。

(a) 原画像 (の一部)　(b) depth1 を仮定して変換

図 4: 実験に用いた画像とその変換例

がうまくまとまらない場合が多くなるが、拘束をかけることによって、それをかなりの割合で回避できる。

これらの結果から、動きを小さくすること、つまり、フレーム間隔を小さくするのが精度を良くするために有効であり、拘束を付加することには大幅な失敗を減らす効果が大きいと言える。

5.2　注目シーンによる要約と評価

注目シーンの抽出

実際のシーンを用いて、注目シーンによる映像要約を作成し、それを用いて評価を行った。行動内容は作業、料理、デスクワークなどである。 動き推定は、4 フレームごとに行い、停留シーン検出における画像中心の動きベクトルの閾値は 240 画素とした。

要約生成例を図 7 に示す。左一列に停留シーンが並べられている。縦方向が時間の経過を表す。この実験例では、部屋からキッチンへの移動や、流し台での作業とコンロでの作業の間での移動等の大きな動きを伴ったシーンの前後が区別されている。

また、右側の黒い縁のついた画像群が同一区間に起こった積極的な注目シーンである。この映像からは、材料の野菜の調理シーン (6~19) や調味料を加えているシーン (22,26,27,33) が抽出されている。また、背景の 1 部などを検出してしまう誤検出は 3 箇所 (4,43,44)

(a) 拘束なし　(b) 拘束あり

図 5: 動き推定の結果: 拘束条件の付加が有効に働く例

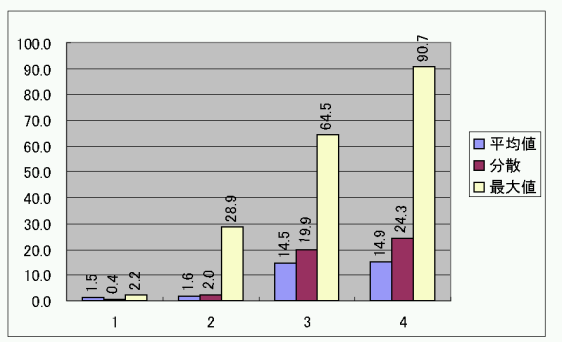


図 6: 動き推定の定量的評価: 1: 動き小で拘束あり、2: 動き小なし、3:動き大で拘束あり、4:動き大で拘束なし

表 1: 本システムのビューアとビデオデッキを用いた情報検索の効率の比較 (単位は秒)

					平均
本研究の結果を利用	315	380	405	360	365
ビデオデッキを使用	424	635	713	800	643

であった。

材料をカットするシーンがやや冗長であるが、図 7 を見ることによって、作業の概略を知ることができる。以上のように、概ね良好な結果が得られることがわかってきたが、以下ではより詳しい評価を行う。

解答時間による評価

4.2 節で述べたように、解答時間による評価を行った。映像としては、上記の料理、デスクワーク、作業を含む、約 34 分の映像を用いた。被験者は 8 人とし、4 人ずつの 2 グループに分けた。また、質問は 4 問とし、映像によって細かい作業内容を再確認する必要がある内容とした。

解答にかかった時間を表 1 に示す。この結果により、本システムを用いた行動映像へのアクセスは、ビデオテープを用いた方法に比べ、1.5 倍 ~ 2 倍程度の効率が得られることが推定できる。

ただし、サンプルの数が少ないため、今後十分な数の実験を行うことが課題になっている。

主観的評価

注目シーンとして選ばれた 46 枚、一定の間隔で選びだした 46 枚、ランダムに選びだした 92 枚のフレーム (画像) を合わせてフレームの母集合を作り、この中から被験者に 46 枚のフレームを選んでもらった。被験者は 10 人である。

実験により得られた、合致率の平均を表 2 に示す。平均的要約と各人の要約との合致率が 74% であることから、個人差を考えない場合には、7 割強の一致を目標にすれば良いことがわかる。

到達率の平均を表 3 に示す。注目シーン検出による到達率は、平均 63% に達し、等間隔選択の平均 46%

図 7: 注目シーンによる要約例: 前半では、本を見るという冗長な場面は簡潔にまとめられ、後半では、作業内容の詳細が提示されている。

と比べ、明らかに有効であることが言える。

また、平均的要約の中に含まれ、注目シーンとして選ばれなかったものにも、動きの特徴などから注目シーンとなり得るものが 7 枚あった。注目対象検出の改良により検出が可能になったとすると、到達率は 80% 程度まで上昇すると考えられる。

ただし、注目シーン検出の考え方では抽出できないフレームも数枚あった。重要な物体がシーン中に静止しており、それを短時間見たものである。作業時には特に注意して見なかったことが推測される。この様

表 2: 合致率の平均値

	平均的要約	注目シーン	等間隔選択
合致率の平均	68%	50%	38%

表 3: 到達率の平均値: ※ は全ての注目シーンが抽出できるような理想的な場合を表す

	注目シーン	等間隔選択	注目シーン※
到達率	63%	46%	77%

なシーンについては、本研究の方法では抽出できず、物体認識等の併用が必要になる。

6 まとめ

本稿では、頭部に装着したカメラにより撮影された映像を、記憶の補助や経験の共有に使うための要素技術について提案した。具体的には、利用者がじっと同じ場所付近を見続けた場面や何かをよく見ようとした場面を注目シーンとして抽出し、映像を要約する方法について提案した。さらに、実際の行動記録映像を構造化、要約し、その評価を行なった。

今後、様々な映像内容に関して十分な実験を行い、有効性を確認していく予定である。

参考文献

[1] 飯島俊匡, 石上陽一, 川嶋稔夫, 青木由直: “日常生活映像から検出された人物像によるエピソード想起”, 信学技法, PRMU97-196

[2] 石川陽一, 飯島俊匡, 川嶋稔夫, 青木由直: “エピソード映像の時空間的階層呈示による記憶想起”, 信学技法, PRMU98-186

[3] Jebara,T., Schiele,B., Oliver,N., Pentland,A., “DyPERS: Dynamic Personal Enhanced Reality System”, MIT Media Laboratory, Perceptual Computing Technical Report ♯463

[4] 大出純哉, 中村裕一, 大田友一: “ビデオ映像による個人行動記録システムにおける注目シーンの検出”, 第 5 回画像センシングシンポジウム, pp.179-184, 1999.

[5] 大出純哉, 中村裕一, 大田友一: “ビデオ映像による個人行動記録・要約システム -注目シーン検出とその安定化- ”, 第 5 回知能情報メディアシンポジウム, pp.83-90, 1999.

[6] Szeliski,R. and Shum,H.: “Creating Full View Panoramic Image Mosaics and Environment Maps”, Proc. SIGRAPH, pp.251-258, 1997.

[7] Bergen,J., Anandan,P., and Hanna,K.: “Hieratical model-based motion estimation” Proc. ECCV, pp.237-252, 1997.

[8] Smith,M. and Kanade,T.: “Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques” IEEE CVPR,1997

[9] Margery Eldridge, Michael Lamming & Mike Flynn.: “Does a Video Diary Help Recall” People and Computers VII,Cambridge University Press, Cambridge, 1992, pp257-269