様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　３１　日現在

機関番号：　１４６０３
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３０１１５

研究課題名（和文）撮影者の意図に起因する映像中の重要領域の推定

研究課題名（英文）Inferring important regions attributed to a videographer's intention

研究代表者
中島　悠太（Nakashima, Yuta）

奈良先端科学技術大学院大学・情報科学研究科・助教

研究者番号：７０６３３５５１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：映像中の注目領域推定は様々なアプリケーションでの応用が考えられる技術であり、現在までに多くの手法が提案されている。その多くは生物の持つ視覚システムをモデル化したものであり、特に映像視聴時の注目領域とは異なるものであると考えられる。本研究では、撮影者の意図に着目することで、映像視聴時の注目領域を推定する手法を提案する。また、この手法を応用可能なアプリケーションとして、人物の自動プライバシー保護処理手法と映像要約手法を開発し、それぞれの有用性を示す。

研究成果の概要（英文）：Inferring important region in a video is a technique that can be used for various applications, and a number of approaches have been proposed so far. Most of these approaches model the visual system of animals and inferred regions may not be very consistent with expected important regions, especially when watching a video. In this project, we focus on a videographer’s intention and propose an approach for important region inference when watching a video. We also develop fully-automatic video privacy protection and video summarization as potential applications of the proposed important region inference approach and experimentally demonstrate their performances.

研究分野：　コンピュータビジョン・パターン認識

キーワード：　重要領域推定　映像　プライバシー保護　映像要約

１．研究開始当初の背景

視覚的注意モデルは画像・映像中の重要領域を推定のための手法で、映像・画像のタスクに依存せず、視聴者が注目する領域（注目領域）を推定する。このモデルは映像・画像の自動要約や圧縮、デバイス適応など様々な応用が可能なことから、生物の視覚システムが持つ生物学的な特徴をモデル化した Visual Saliency [Itti 1998]、映像中で予測できないような変化をする領域を検出する Bayesian Surprise [Itti 2005]、人間の顔に注意が向きやすいという性質を考慮して上記のモデルなどに顔検出を援用するモデル [Ma 2005] など、様々な手法が提案されている。しかし、特に映像においては、例えば図 1(a) のフレームに対して、視聴者は中央付近の人物が重要領域と考えると予想される一方で、例えば [Itti 1998] のモデルでは図 1(b) で示すようなコントラストの高い領域を注目領域として推定しており、従来の視覚的注意モデルでは期待する結果が得られないという問題があった。

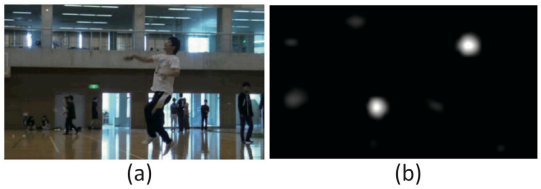

図 1　(a)映像のフレーム例と(b)視覚的注意モデルの推定結果

２．研究の目的

本研究では、上記の問題点を解決した新しい注目領域推定のアプローチとして、撮影者の意図に着目した手法を提案する。この手法では、映像シーン中のオブジェクト（人・ものなど）の動きに対して撮影者がどのようにカメラを動かしたかに撮影者の意図が反映されるという考えのもと、注目領域推定の新たなモデルを構築する。加えて、既存の視覚的注意モデルや、実際の人間の注視と比較することにより提案手法の有効性を確認する。

さらに、提案手法に関連するアプリケーションとして、人物の自動プライバシー保護処理手法と映像の自動要約手法を開発し、本研究で提案する注目領域推定手法の応用可能性を示す。

３．研究の方法

（１）撮影者の意図に基づく注目領域推定

前述のように、現在までに提案されている注目領域推定手法には、生物の視覚的注意モデルを利用するものが多い。これは、コントラストの高い領域や動きのある領域に視覚的注意が向くという生物の視覚の性質をモデル化したものである。しかし、特に映像を視聴する際においては、視聴者はその映像で表現される内容を理解するために必要な領域に対して選択的に注目すると考えられる。

本研究では、特に一般ユーザが撮影した未編集の映像においては、このような視聴者の視覚的注意が撮影者の意図に誘発されるものであると考えた。これは、映像を撮影する際に、撮影者は視聴者に見せたいものに対して、中央に据える、追いかけるなどのようにカメラを動かすと考えられ、視聴者はこのようなカメラの動きから、その映像を理解するために必要な領域を決定し、その領域に注目するという仮定に基づくものである。このような撮影者の意図は、シーン中のオブジェクトの動きと撮影者のカメラの動きに反映されるものと考えられる。シーン中のオブジェクトを追跡することによって、その動きとカメラの動きを合わせた軌跡が得られることから、このような軌跡を特徴量として利用することにより、注目領域の推定が可能であると考えられるが、一般に、任意のオブジェクトの追跡は困難である。そこで本研究では、オブジェクトの追跡に替えて、映像中の任意の点を追跡することで点軌跡を抽出する手法 [Sand 2006] を利用する。

本研究では各点軌跡それぞれに対して注目・非注目領域を判定する識別器を学習し、さらに(i)同一オブジェクト上の点は同じ注目・非注目領域に含まれる可能性が高い、(ii)注目・非注目領域は時間的に連続する、の２点を図 2 に示すような Markov Random Field（MRF）を利用してモデル化した。

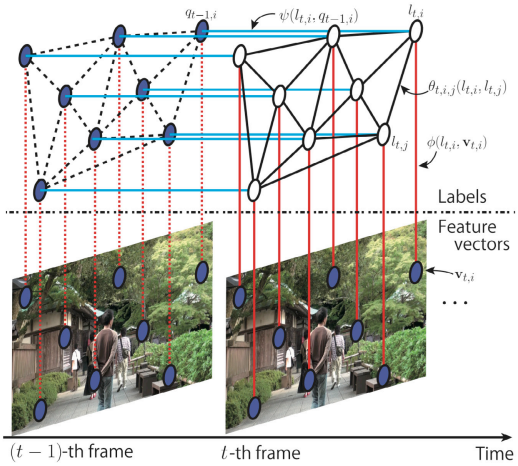

図 2　MRF による注目・非注目領域モデル

（２）人物の自動プライバシー保護処理

ソーシャルネットワーキングサービス（Social Networking Service; SNS）やカメラ付きデバイスの普及により、誰もが画像・映像を撮影・公開できるようになった。既存のプライバシー保護処理手法では、例えば [Dufaux 2008] のように全ての人物に対してプライバシー保護を適用するものが多いが、上記のような状況においては、①全ての人物に対してプライバシー保護を適用すると映像が無意味なものになる可能性がある、②既存のぼかしや塗りつぶしなどの保護処理方法では、保護対象の表情などが失われる、などの問題があった。

そこで本研究では、①に対して、注目領域

推定手法をベースとして人物に特化した重要人物推定手法を開発し、この手法に基づいたプライバシー保護処理を提案した。この手法では、人物検出・追跡により得られた人物の軌跡をもとに、（1）の注目領域推定と同様のモデルにより、人物間の空間的関係と重要・非重要人物の時間的連続性を考慮して重要人物と非重要人物を識別し、非重要人物についてはその映像にとって不要であるとして背景推定により除去する（図 3）。

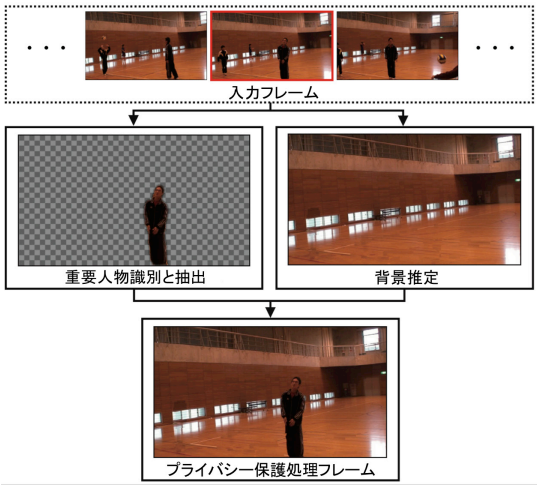

図 3　提案するプライバシー保護処理手法の全体図

また②については、近年提案された画像処理手法である Image Melding [Darabi 2012] を用いて第三者の顔を保護対象画像に混合することで、人物の表情を維持可能かつ視覚的に自然な保護処理手法を開発した（図 4）。

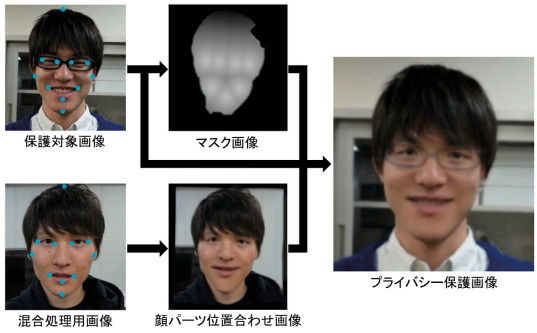

図 4　表情を維持した視覚的に自然なプライバシー保護処理手法の全体図

（3）テキストに基づく映像要約生成

映像要約は、長時間にわたる冗長な映像からコンパクトな映像を生成する技術であり、現在までに様々な手法が提案されている。多くの手法は、スポーツやニュースの映像を対象とし、スポーツのルールや編集方法に関する事前知識を用いるもの、もしくは映像セグメント（映像を短く区切ったもの）を、色ヒストグラムなどの低レベル特徴量を利用してクラスタリングし、各クラスタの代表セグメントを要約として出力するなど、要約映像の内容を指定することができないものであった。

本研究では、近年多くのウェブサイトで見られるビデオブログ（図 5、映像とその映像に対応するテキストで構成されるブログ）の映像を、長時間撮影された記録映像に対して映像要約を適用することにより生成する手法を提案する。この手法では、ビデオブログのテキストに着目し、テキストに対応する映像セグメントを抽出することで、要約映像の内容を直接コントロール可能としている点で上記既存手法とは異なるものである。

図 5　ビデオブログ

具体的には、テキストから名詞を抽出するとともに、映像セグメント中からオブジェクトを検出し、それらの一致度合いによってテキストと映像セグメントの類似度を定義した。この類似度に基づいて定義した目的関数を最適化することにより、要約映像に含まれる映像セグメントを選択する（図 6）。この類似度算出の際に、（1）の注目領域推定手法を利用することにより、撮影者の意図に基づいた要約映像の生成が可能であると考える。

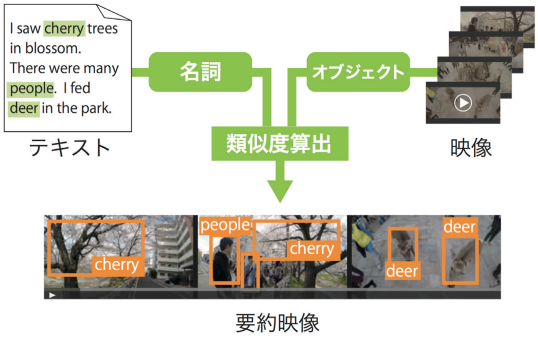

図 6　映像要約手法の全体図

4．研究成果

（1）撮影者の意図に基づく注目領域推定

図 7 に元の映像フレーム、実際に撮影者によって付与された注目領域、本研究で提案した注目領域推定手法の推定結果、及び視覚的注意モデルによる比較手法 [Hou 2008] による推定結果を示す。比較手法では、コントラストの強い領域が推定されている一方、提案手法では実際に付与された注目領域に近い推定結果が得られていることがわかる。図 8 に提案手法、比較手法（[Hou 2008]、[Harel 2006]、[Itti 1998]）、実際の人間による注目領域（視線計測装置を利用して取得）が、撮

影者によって指定された注目領域とどの程度一致するかを Area Under Curve（AUC）により評価した結果を示す。これより、提案手法は他の評価結果に比較して高い性能を示すことが明らかとなった。実際の人間の注目領域が低い値となっているか、これは人間の注目領域が点（注視点）によって与えられるために、撮影者によって指定された注目領域を被覆できなかったためである。

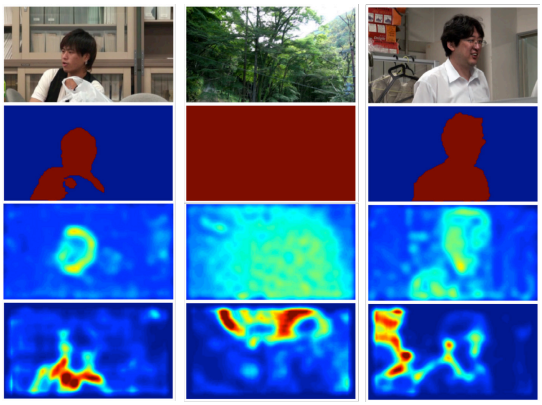

図 7　映像フレーム例（1 列目）、撮影者が指定した注目領域（2 列目）、提案手法による注目領域の推定結果（3 列目）、[Hou 2008] の出力（4 列目）

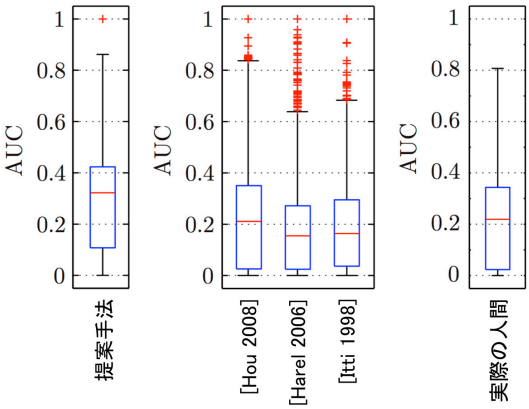

図 8　AUC による注目領域推定の評価と比較

（2）人物の自動プライバシー保護処理

①について、図 9(左)に時間的連続性の考慮の度合い（α）に対する重要人物・非重要人物の識別精度を示す。時間的連続性をほぼ考慮しない（α＝0）付近で AUC が最大となることが明らかになった。これは、識別の際に人物の追跡結果を利用しているために、すでに時間的な連続性を利用した識別となっているため、さらに時間的連続性を考慮する必要がないことを示すものであると考えられる。図 9(右)は(i)提案手法と(ii)既存手法[Nakashima 2011]、(iii)空間的な関係、及び時間的連続性を利用せずにサポートベクタマシン（SVM）のみを識別器として用いる手法、(iv)人間による識別の性能を Receiver Operation Characteristics (ROC) 曲線によって評価した結果を示す。提案手法では人間による識別には及ばないものの、その他の手法よりも高い性能が得られることが明らかになった。

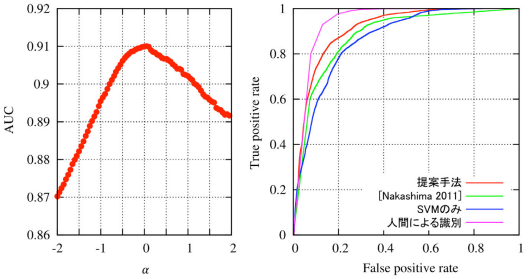

図 9　時間的連続性の考慮度合いと識別性能の AUC による評価（左）と、既存手法などとの ROC による比較（右）

②について、図 10 に提案した手法によるプライバシー保護映像を示す（4 段目）。図中 1 段目の赤枠を重要人物、青枠は非重要人物であるものとし、非重要人物に対してプライバシー保護処理を適用した。比較手法として、ぼかし（2 段目）、モーフィング（3 段目）について、提案手法と合わせてプライバシー保護処理画像が視覚的に自然かを 5 段階で評価したところ、ぼかし 2.0、モーフィング 3.9、提案手法 4.2 であった。これにより、提案手法で視覚的に自然なプライバシー保護処理画像が生成可能であることを示した。

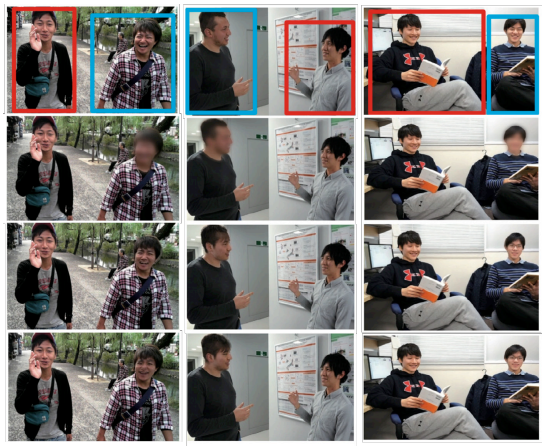

図 10　画像例（1 段目）、ぼかし（2 段目）、モーフィング（3 段目）、提案手法（4 段目）

（3）テキストに基づく映像要約生成

提案手法の評価のために、3 種のテキストを利用して単一の映像群の要約を生成した。比較のために、(a)一定間隔でのサンプリング、(b)クラスタリングに基づく手法、(c-i)～(c-iii)テキスト 1 による提案手法、(d-i)～(d-iii)テキスト 2 による提案手法、(e-i)～(e-iii)テキスト 3 による提案手法について評価を実施した。テキスト 1 ～ 3 について、(i)は提案手法であり、(ii)と(iii)はそれぞれ提案手法の一部の機能を除去したものである。これらの要約映像に対して、被験者がテキストと映像を閲覧し、そのテキストがブログの本文であるとした場合にどの映像がふさわしいかをそれぞれ 1 点から 5 点で評価した。テキスト 1 ～ 3 が

与えられたときのそれぞれの映像の評価結果を図 11(上)～(下)にそれぞれ示す。図中、緑色で示す結果が高いほど、提案手法の性能が高いことを表す。この結果から、テキスト１についてはクラスタリングに基づく手法(b)が最も高い評価を得たものの、テキスト２、及びテキスト３については提案手法が良い評価を得た。テキスト１については、テキスト自体が映像群全体の内容に近いものであったために、より多くの内容を含む(a)、及び(b)で評価が高くなったと考えられる。これより、ビデオブログの映像生成において提案手法が望ましい性質を持つことが示せた。

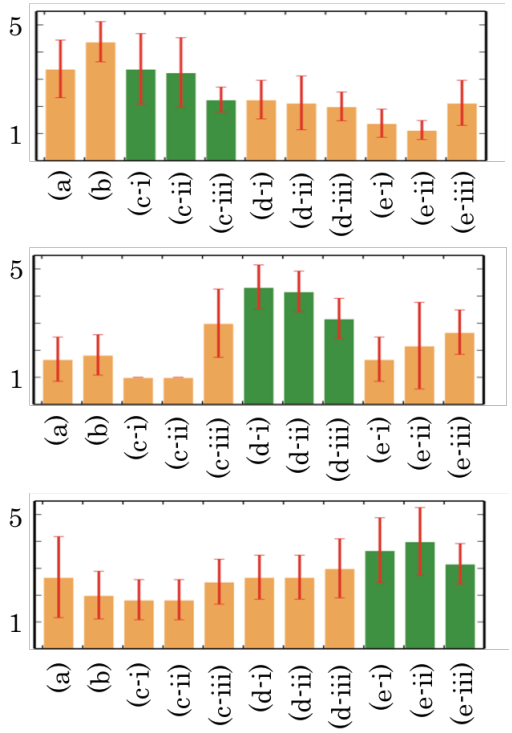

図 11　生成された要約映像の評価結果

＜引用文献＞

[Itti 1998] L. Itti et al., “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. PAMI Vol. 20, No. 11, pp.1254–1259, 1998

[Itti 2005] L. Itti and P. F. Baldi, “A principled approach to detecting surprising events in video,” Proc. CVPR, pp.631–637, 2005

[Ma 2005] Y. F. Ma et al., “A generic framework of user attention model and its application in video summarization,” IEEE Trans. Multimedia, Vol.7, No.5, pp.907–919, 2005

[Sand 2006] P. Sand and S. Teller, “Particle video: long-range motion estimation using point trajectories,” in Proc. CVPR, pp.2195–2202, 2006

[Dufaux 2008] F. Dufaux and T. Ebrahimi, “Scrambling for privacy Protection in video surveillance systems,” IEEE Trans. CSVT, Vol.18, No.8, pp.1168–1174, 2008

[Darabi 2012] S. Darabi et al., “Image melding: Combining inconsistent image using patch-based synthesis,” ACM ToG, Vol.31, No.2, pp.82:1–82:10, 2012

[Hou 2008] X. Hou and L. Zhang, “Dynamic visual attention: Searching for coding length increments,” Proc. NIPS, pp.681–688, 2008

[Harel 2006] Harel et al., “Graph-based visual saliency,” Proc. NIPS, pp.545–552, 2006

[Nakashima 2011] Y. Nakashima et al., “Intended human object detection for automatically protecting privacy in mobile video surveillance,” Multimedia Systems, Vol.18, No.2, pp.157–173, 2012

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　3　件）

[1] Y. Nakashima, N. Babaguchi, and J. Fan, “Privacy protection for social video via background estimation and CRF-based videographer’s intention modeling,” IEICE Trans. Information and Systems, Vol.E99-D, No.4, pp.1221–1233, Apr. 2016 (DOI: 10.1587/transinf.2015EDP7378) (査読有)

[2] Y. Nakashima, T. Ikeno, and N. Babaguchi, “Evaluating protection capability for visual privacy information,” IEEE Security & Privacy, Vol.14, No.11, pp. 55–61, Feb. 2016 (DOI: 10.1109/MSP.2016.3) (査読有)

[3] N. Babaguchi and Y. Nakashima, “Protection and utilization of privacy information via sensing,” IEICE Trans. Information and Systems, Vol. E98-D, No. 1, pp. 2–9, Jan. 2015 (DOI: 10.1587/transinf.2014MUI0001) (査読有)

〔学会発表〕（計　7　件）

[4] Y. Nakashima, “Point trajectory-based inference of what the videographer wanted to capture,” 画像の認識・理解シンポジウム(MIRU)講演論文集, SS1-13, Jul. 2015, ホテル阪急エキスポパーク(大阪・吹田) (査読無)

[5] Y. Nakashima, “Facial expression preserving privacy protection using image melding,” Proc. IEEE Int. Conf. Multimedia and Expo, 6 pages, Jun. 2015, トリノ(イタリア) (査読有)

[6] 大谷 まゆ, “テキストと映像の類似度を用いた映像要約,” 電子情報通信学会 技術研究報告, PRMU2014-95, Jan. 2015, 奈良先端科学技術大学院大学(奈良・生駒)(査読無)

[7] 大谷 まゆ, “テキスト記述を用いてユーザ意図を反映する映像要約,” 電気関係学会

関西連合大会講演論文集, pp.384-385, Nov. 2014, 奈良先端科学技術大学院大学（奈良・生駒）(査読無)
[8] 小山 達也, 中島 悠太, 馬場口 登: "画像のコンテキストを保持した視覚的に自然なプライバシー保護処理," 電子情報通信学会技術研究報告, PRMU2013-205, Mar. 2014, 早稲田大学(東京) (査読無)
[9] Y. Nakashima and N. Yokoya, “Inferring what the videographer wanted to capture,” Proc. 2013 Int. Conf. Image Processing, pp.191–195, Sept. 2013, メルボルン(オーストラリア) (査読有)
[10] T. Koyama, Y. Nakashima, and N. Babaguchi, “Real-time privacy protection system for social videos using intentionally-captured persons detection,” Proc. 2013 Int. Conf. Multimedia and Expo, 6 pages, Jul. 2013, サンノゼ(アメリカ) (査読有)

〔その他〕
ホームページ等
http://www.n-yuta.net/kakenhi-project-2013-2015/

６．研究組織
(1)研究代表者
中島　悠太（NAKASHIMA, Yuta）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：70633551