

静止画内オブジェクトへの指示による動画検索手法の提案

川手裕太[†] 岡部誠^{†,*} 尾内理紀夫[†] 平野廣美[‡] 三條正裕[‡]

[†]電気通信大学 〒182-8585 東京都調布市調布ヶ丘 1-5-1

[‡]楽天株式会社 楽天技術研究所 〒140-0002 東京都品川区東品川 4-13-9

*科学技術振興機構 さきがけ

E-mail: [†]kawate@onailab.com, m.o@acm.org, onai@cs.uec.ac.jp

[‡]{hiromi.hirano, masahiro.sanjo}@mail.rakuten.com

あらまし 動画共有サイトの普及により、ユーザは多数の動画の視聴が可能になった。ユーザは動画に付与されたテキストタグを用いて、視聴したい動画を検索している。しかし、例えば「左向きの車が右から左に動いている」といった動画内オブジェクトの向き・動きを検索することは、テキストタグを用いた従来の手法では困難なことが多くある。本論文では、静止画内オブジェクトの向きなどを直接指示操作することにより動きを含む類似動画を検索する手法を提案する。それを実現するため SIFT(Scale-Invariant Feature Transform)と Particle Video を用いた。SIFT は入力静止画内オブジェクトを動画から検索するために使用し、Particle Video は動画内オブジェクトの軌跡を抽出するために使用した。

キーワード 動画検索, SIFT, Particle Video

1. はじめに

動画共有サイトの普及により、ユーザは数多くの動画を視聴できる環境にある。しかし、ユーザの目的の動画を見つけ出すことが困難なことが多くある。例えば、ユーザの目的の動画が「車が右から左に動く」という動画であった場合、この「車が右から左に動く」というテキストのクエリを入力しても思い通りの結果が出てこない。この原因は、動画の検索手法として動画のタグを用いた検索を行っているからである。タグはテキストでユーザの主観によって動画の内容に関係ある単語が付けられることが多い。そのため、「車が右から左に動く」という様子の動画に対して、ユーザがその動画にタグを付けるとするとタグは、「車」という動画内に登場しているオブジェクトの単語であり、「車が右から左に動く」という動画内の内容を表現しているテキストではない。よって、ユーザが「車が右から左に動く」という動画を見つけ出したいときは「車」というテキストで動画を検索し、一つ一つ動画を確認する必要がある。ユーザが一つ一つ確認するという点では動画検索の体をなしていない。本論文において、テキスト情報を用いない動画内のオブジェクトの向きや動きを検索することが可能なシステムを提案する。

本論文の構成は次のとおりである。2 章で関連研究について述べ、既存の研究の問題点、本論文との差異

を述べる。3 章で提案システムの概要を述べ、4 章で使用する特徴量について述べ、5 章で提案システムを用いた実験について述べ、6 章で本論文のまとめを述べる。

2. 関連研究

テキストを用いない動画検索に関する研究では様々な手法が提案されている。

福田ら[1]の手法は、取得したい動画と同じような動画をクエリとして与える。その動画から静的・動的特徴量を抽出し、これを特徴量データベースと比較を行う。この手法の問題点は動画を用意しないといけないという点である。例えば、「車が右から左に動く」という様子の動画を取得したい時には、予め「車が右から左に動く」様子の動画を用意しないといけない。一方、本論文で提案する手法は取得したいオブジェクトの静止画を 1 枚用意し矢印などを入力するだけである。この点が福田らの手法と大きく異なる点である。

有村ら[6]の手法は、取得したい動画の動きベクトル列、動きベクトルの経過時間列、形状データ列をクエリとして与える。それらの特徴量データベースと比較する。この手法の問題点はオブジェクトの検索を形状データ列で比較する点である。形状データ列はフレー

ム毎に一定の大きさに分割したブロックがそのフレーム内での平均的な動きベクトル列に類似しているか否かという特徴量である。つまり、オブジェクトが異なるものであっても形が類似していたらマッチングする。一方、本論文で提案する手法は取得したいオブジェクトと特徴量データベースとを SIFT を使用し比較を行うため、有村らの手法より良い精度が得られる。

また、動画内オブジェクトの直接操作の研究として、Goldman ら[2]の手法がある。この手法は、動画内のオブジェクトをドラッグ&ドロップで直接操作を行っている。しかし、この手法では直接操作できる範囲が連続した動画内ではない。一方、本論文の提案する手法では、検索可能な範囲が連続していない動画も含まれている。この点が Goldman らの手法とは大きく異なる点である。

3. システム概要

ユーザが本システムを使用して、動画を検索する具体的な手順は以下のようになる。

0. 前準備として、多数の動画をシステムに入力して、SIFT や Particle Video といった特徴量のデータベースを作成する。
1. ユーザは検索したいオブジェクトを含む静止画を入力としてシステムに与える。
2. ユーザは入力された静止画の上に、マウสดラッグといったインタラクティブな操作で検索したいオブジェクトの向きを矢印で入力する。
3. システムが入力された静止画・矢印と特徴量データベースとを比較する。結果として類似度が高い順に動画内のフレームと軌跡を出力する。ただし、候補が多い場合は4へ。フレームの静止画を選択するとそれに対応する動画が出力される。
4. ユーザは検索したいオブジェクトの動きを矢印で入力する。
5. システムが入力された矢印と特徴量データベースとを比較する。結果として類似度が高い順に動画内のフレームと軌跡を出力する。

このようにして、容易かつ直感的な動画検索を可能とする。

オブジェクトの向きを変更させたい場合の入力方法は図1の方法である。検索したいオブジェクトが含まれた静止画を与える(1a)。オブジェクトの前方の部分が右側に(1b)、オブジェクトの後方の部分が左側に(1c)と、オブジェクトの向きをどのように変更したいかを矢印入力で指定する。結果として指定したオブジ

ェクトの向きが変わった静止画が出力される(1d)。

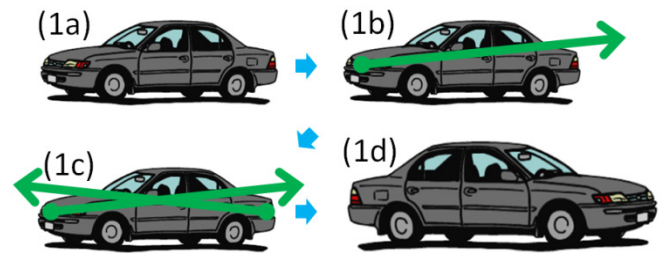


図1:オブジェクトの向き変更の入力方法とその結果

オブジェクトの動きを指定したい場合の入力方法は図2の方法である。検索したいオブジェクトが含まれた静止画を与える(2a)。オブジェクトを移動させたい方向へ矢印を入力する(2b)。結果として車が移動する動画が検索され、出力される(2c)。

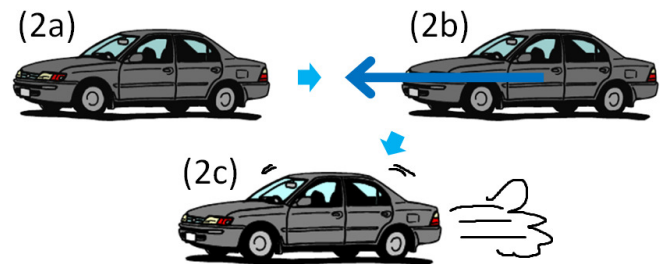


図2:オブジェクトの動き指定の入力方法とその結果

4. 比較に使用する特徴量

動画の検索のために、入力された静止画・矢印と特徴量データベースとを比較する。その比較に使用する特徴量として2種類の特徴量を使用する。1つ目は静止画の局所特徴を検出し記述する SIFT[3]である。もう1つは動画の各フレームとその後のフレームの内の各ピクセルのオプティカルフローを行った結果を用い、各フレームに配置した多数の点の軌跡を抽出する Particle Video[7]である。

前者の SIFT を使った類似静止画検索の手法は既に存在する。それらの手法の中で使われている特徴量のマッチング[4]を行った。その結果、入力された静止画内のオブジェクトと同じ横向きであれば多くのマッチングが抽出できる(図3左)。ただし、オブジェクトが横向きから正面へと向きを変えただけでそのマッチングは上手く働かなくなる(図3右)。

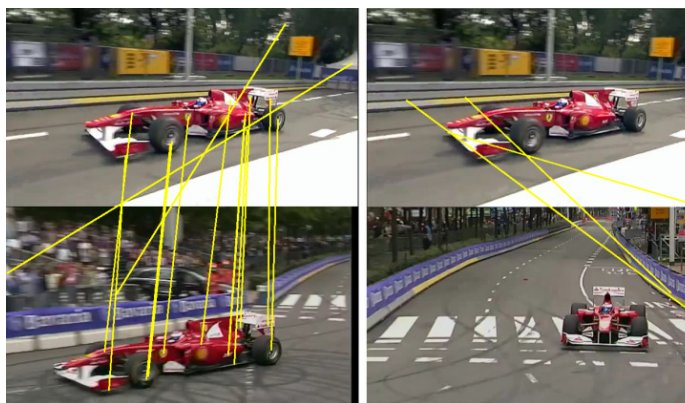


図 3:SIFT を用いたマッチング

ここで、動画を再生すると動画内のオブジェクトの向きが変わり、オブジェクトの向きが入力された静止画内のオブジェクトと同じ向きとなるフレームが登場することがある。そのフレームのオブジェクトは入力された静止画内のオブジェクトと SIFT のマッチングが多く抽出できるため、入力された静止画内のオブジェクトと同じオブジェクトとして検索することができる。ここで、Particle Video を用いると連続する動画であれば入力された静止画内のオブジェクトと同じ向きのオブジェクトと異なる向きのオブジェクトが同じオブジェクトであると判定できる(図 4)。これにより、入力された静止画内のオブジェクトと向きが異なったオブジェクトも検索することが可能になる(図 5)。

また、Particle Video は動画内に多数配置した点の軌跡を抽出するので、動画内のオブジェクトの軌跡を抽出することができる。これをオブジェクトの軌跡の検索にも使用する。



図 4:SIFT と Particle Video を個々で用いた静止画検索



図 5:SIFT と Particle Video を同時に用いた静止画検索

5. 実験

5.1. 実験方法

実験では 3 章で述べた前準備をまず行った。特徴量のデータベースを作成する動画として、YouTube に投稿されている動画の中から 3 種類のオブジェクト(車・飛行機・馬)が含まれている動画をシーンごとに分割し、種類ごとに 20 本の動画(640×360 ピクセル)を用意した。入力静止画は Google 画像検索によって得られた静止画を使用した。実験に使用した計算機の CPU は Intel Core i7-860 2.80GHz、メモリは 4.0GB である。

5.2. 実験結果

表 1 はそれぞれのデータベースの動画の合計フレーム数と、その検索にかかった時間である。

オブジェクト	車	飛行機	馬
合計フレーム数	3718	3741	4450
検索時間(s)	4.43	3.10	4.04

表 1: データベースの合計フレーム数と検索時間

車に対する検索結果が図 6 である。図 6 左では図 1 で表したオブジェクト向きを変更させたい場合に入力する緑矢印の 1 本目を車の後方の箇所が右側へ、2 本目を車の前方の箇所を左側へと入力した。つまり、「左向きの車」というクエリを矢印入力によって表現した。次に、図 6 右では緑矢印の 1 本目は車の後方の箇所が上側へ、2 本目は車の前方の箇所を下側へと入力した。その後、2 で表したオブジェクトの動きを指定したい場合に入力する青矢印を右斜め上から左斜め下へと入力した。つまり、「手前向きの車が左下へと動く」というクエリを矢印入力によって表現した。結果は GUI で対応するフレームを類似する順番に 3 枚表示させた。また、表示されているフレームを選択するとそれに対応する動画がそのフレームから再生される。

飛行機に対して同様に入力を行った検索結果が図 7、馬に対して同様に入力を行った検索結果が図 8 である。

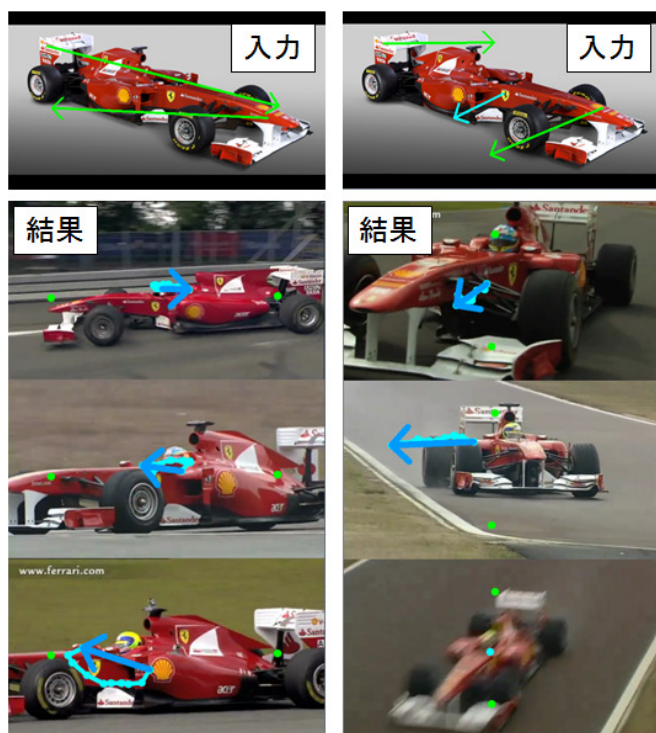


図 6: 実験結果 1 (車)

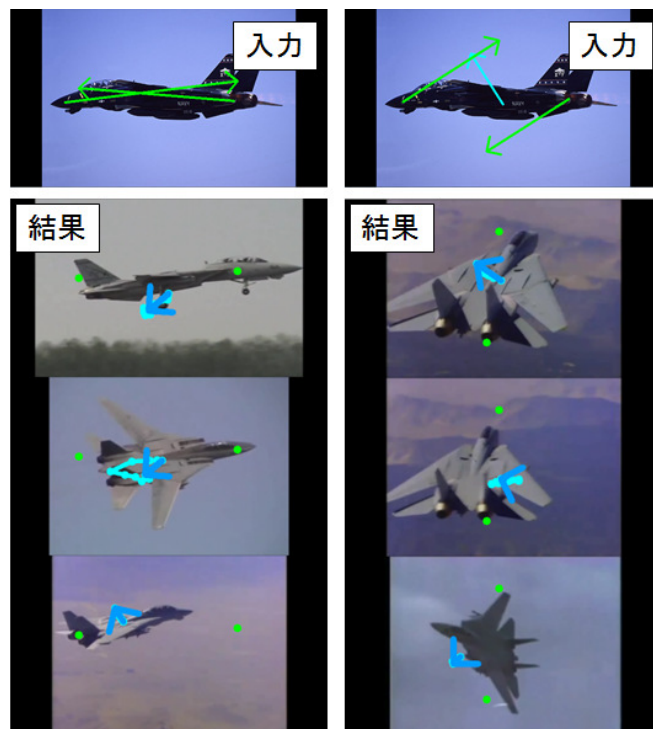


図 7: 実験結果 2 (飛行機)



図 8: 実験結果 3 (馬)

6. まとめ

本論文では、静止画内オブジェクトの向きなどを直接指示操作することにより動きを含む類似動画を検索する手法を提案した。実験において、入力静止画・矢印によって動画内のオブジェクトの向きを検索できることを確認した。

今後の課題としては、まず Particle Video という特徴量は抽出する時間が 1 フレームあたり約 7 秒と非常に時間がかかる。これを解決するためには、オブジェクトの追跡について別の方法を検討する必要がある。

次に、SIFT のマッチングにおいて色情報を使用していない。よって、例えば、車のデータベースを大量に用意した場合、形は同じだが色の違う車が検索結果として表示される可能性がある。この場合の対処について、色の違う車を許容して検索結果を出すのか、色の違いを許さず色情報を使用する方法を考えるのかなどどうさせるかを考える必要がある。

3 つ目に、現段階では入力静止画の SIFT を予め求め、各動画のフレームとのマッチングを予め行っているため、入力静止画を新たに用意してもすぐに検索を行うことができない。この原因は、各フレームのオブジェクトの切り出しを行っていないので、ノイズが多く出現し、入力静止画に一番似ているフレームの検出が困難であるからである。また、入力静止画と各動画のフレームとを 1 対 1 対応でマッチングを行っており、非常に時間がかかるという問題点もある。前者の問題点は切絵[8]などをオブジェクトの切り抜きが行える手法を使用すれば、ノイズを減らすことが可能になる。後者の問題点は、各動画から抽出した SIFT を同様な集合にまとめた visual dictionary[5]を作成し、使用すれば SIFT マッチングの高速化が実現できる。

4 つ目に、カメラが固定されていない動画では動画内のオブジェクトの動きが実際の動きと一致しないことがある。例えば、動画内のオブジェクトが実際に右から左に動いている時にカメラがオブジェクトを右から左へと追っかけた場合、撮影された動画内でオブジェクトの位置はほとんど変わらない。この問題を解決するためには背景の動きを追いかけるという方法が考えられる。

最後に、現状の緑矢印 2 本だけを使用したオブジェクトの向きの変更では、手前向きと奥向きとの判別がつきにくいという問題点がある。これを解決するために新たに前後を指定する矢印を導入するなどインターフェースの改良を行っていきたいと考える。

参 考 文 献

- [1] 福田淳、苗村昌秀、和泉吉則、伊藤泰雅、“動画画像検索装置”、特開 2001-134589、2001.
- [2] Dan B Goldman, Chris Gonterman, Brian Curless, David Salesin, Steven M. Seitz. “Video Object Annotation, Navigation, and Composition”, In UIST, pages 3–12, 2008.
- [3] D. G. Lowe, “Object recognition from local scale-invariant features”, Proc. of IEEE International Conference on Computer Vision (ICCV), pp.1150-1157, 1999.
- [4] 藤吉弘亘. Gradient ベースの特徴抽出-SIFT と HOG-. CVIM 160, pp.211-224, 2007.
- [5] Josef Sivic and Andrew Zisserman. “Video Google: Efficient Visual Search of Videos”, volume 4170 of LNCS, pages 127–144. Springer, 2006.
- [6] 有村耕治、佐々木孝幸、“圧縮動画像の検索方法および装置”、特開平 8-194714、1996.
- [7] P.Sand and S. Teller. “Particle Video: Long-range motion estimation using point trajectories”, IJCV, 80(1): 72-91, 2008.
- [8] 清野達也、林貴宏、尾内理紀夫、三條正裕、森正弥、“改良領域拡張法による高速画像切り抜き手法の提案と評価”、情報処理学会論文誌, Vol.50, No.12, pp.3233-3249, 2008.